This collection comprises the posters of 385 out of the 415 papers presented at the European Conference on Computer Vision 2016, Amsterdam, The Netherlands.

All rights related to the posters remain with their authors.

Selected Adobe Shortcuts:

< SHIFT + F >  Search menu
< CTRL + SHIFT + F >  Advanced search menu
< CTRL + 2 >  Zoom in and view the whole page
< CTRL + 3 >  Zoom in and view the page by discarding white boundary

To see the session IDs of posters, please click bookmark button 📋
Oral Sessions
1. Instance Retrieval Challenges

- occlusion
- occlusion changes
- viewpoint changes

2. CNNs and Large Training Datasets

- Pre-trained on ImageNet
  - Internal activations as descriptors
  - Inappropriate data for instance matching
  
  [Gong et al. ECCV'14, Babenko et al. ECCV'15, Kalantidis et al. arXiv'15, Tolias et al. ICLR'16]

- Fine-tuned on landmark classes
  - Internal activations as descriptors
  - Better than ImageNet but still inadequate for instance matching

  [Babenko et al. ECCV'14]

- Weakly supervised fine-tuning
  - Descriptor directly optimized
  - Requires GPS dataset

  [Arandjelovic et al. CVPR'14]

- Our work
  - Training data without any human supervision
  - Lots of training data more appropriate for instance matching
  - Strong automatic annotation for hard negative, hard positive mining, and supervised whitening

3. Training Data without any Human Interaction

- Positive Convolutional Layers
- MAC Layer
- Descriptor

4. CNN Siamese Learning

- Query Convolutional Layers
- MAC Layer
- Descriptor

5. Supervised Whitening

- Whitening: square-root of intraclass covariance matrix: $C_s^{-1/2}$

- Rotation: PCA of interclass covariance matrix in the whitened space: $e_{i,j}C_s^{-1/2}C_dC_s^{-1/2}$

6. Implicit Correspondences

7. Experiments
SSD: Single Shot MultiBox Detector

Wei Liu¹, Dragomir Anguelov², Dumitru Erhan³, Christian Szegedy³, Scott Reed⁴, Cheng-Yang Fu¹, Alexander C. Berg¹
¹UNC Chapel Hill ²Zoox Inc. ³Google Inc. ⁴University of Michigan, Ann-Arbor

OVERVIEW
SSD discretizes bounding boxes space into a set of default box shapes per feature map location, and uses convolution kernel (3 × 3) to predict both the bounding box offsets and object probabilities per location.

COMPARE STATE-OF-THE-ART METHODS

#1: MULTI-SCALE FEATURE MAPS
SSD uses multiple feature maps of decreasing resolution to output bounding boxes of increasing size.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>FPS</th>
<th>batchsize</th>
<th># Boxes</th>
<th>Input res</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN (VGG16)</td>
<td>73.2</td>
<td>7</td>
<td>1</td>
<td>~6800</td>
<td>~1000 × 600</td>
</tr>
<tr>
<td>Fast YOLO</td>
<td>52.7</td>
<td>155</td>
<td>1</td>
<td>98</td>
<td>448 × 448</td>
</tr>
<tr>
<td>YOLO (VGG16)</td>
<td>66.4</td>
<td>21</td>
<td>1</td>
<td>98</td>
<td>448 × 448</td>
</tr>
<tr>
<td>SSD300</td>
<td>74.3</td>
<td>46</td>
<td>1</td>
<td>8732</td>
<td>300 × 300</td>
</tr>
<tr>
<td>SSD512</td>
<td>76.8</td>
<td>19</td>
<td>1</td>
<td>24564</td>
<td>512 × 512</td>
</tr>
<tr>
<td>SSD300</td>
<td>74.3</td>
<td>59</td>
<td>8</td>
<td>8732</td>
<td>300 × 300</td>
</tr>
<tr>
<td>SSD512</td>
<td>76.8</td>
<td>22</td>
<td>8</td>
<td>24564</td>
<td>512 × 512</td>
</tr>
</tbody>
</table>

#2: MORE DEFAULT BOXES
SSD discretizes bounding boxes spaces into many bins, preventing box coordinates averaging when several likely hypotheses are present in the same default box.

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>FPS</th>
<th>batchsize</th>
<th># Boxes</th>
<th>Input res</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSD300</td>
<td>74.3</td>
<td>46</td>
<td>1</td>
<td>8732</td>
<td>300 × 300</td>
</tr>
<tr>
<td>SSD512</td>
<td>76.8</td>
<td>19</td>
<td>1</td>
<td>24564</td>
<td>512 × 512</td>
</tr>
</tbody>
</table>

THE DEVIL IS IN THE DETAILS

1. Data augmentation
2. Ground truth to default box matching
3. Hard negative mining

DETECTION EXAMPLES

REFERENCES

A Recurrent Encoder-Decoder for Sequential Face Alignment

Xi Peng*, Rogerio S. Feris**, Xiaoyu Wang***, Dimitris N. Metaxas*

*Rutgers University  **IBM T. J. Watson  ***Snapchat Research

**Spatial Recurrent Learning

\[ \text{argmin} \sum_{k=1}^{L} \ell(M_k^f, \tilde{f}_{\text{DENC}}(f_{\text{ENC}}(x, \tilde{z}))) \]

\[ \text{argmin} \sum_{k=1}^{L} \ell_{\text{RNN}}(M_k^f, \tilde{f}_{\text{DENC}}(f_{\text{ENC}}(x, \tilde{z}))) \]

**Temporal Recurrent Learning

\[ \text{argmin} \sum_{k=1}^{L} \ell(M_k^f, \tilde{f}_{\text{DENC}}(f_{\text{ENC}}(x, \tilde{z}))) \]

\[ \text{argmin} \sum_{k=1}^{L} \ell_{\text{RNN}}(M_k^f, \tilde{f}_{\text{DENC}}(f_{\text{ENC}}(x, \tilde{z}))) \]

**Architecture

**Experiment


data processing

\[ \text{Identity Loss} \]

\[ \text{Coordinate Loss} \]

\[ \text{Identity Constraint} \]

\[ \text{Constrained Shape Prediction} \]

\[ \text{Supervised Identity Disentangling} \]

\[ \text{Temporal Recurrent Learning} \]

\[ \text{Spatial Recurrent Learning} \]

\[ \text{Encoder-Decoder} \]

**General comparison

7 landmarks

<table>
<thead>
<tr>
<th>TF</th>
<th>FM</th>
<th>300-VW</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMRF</td>
<td>4.43</td>
<td>8.53</td>
</tr>
<tr>
<td>ESR</td>
<td>3.81</td>
<td>7.58</td>
</tr>
<tr>
<td>SDM</td>
<td>4.01</td>
<td>7.49</td>
</tr>
<tr>
<td>IFA</td>
<td>3.45</td>
<td>6.39</td>
</tr>
<tr>
<td>DCN</td>
<td>3.67</td>
<td>6.16</td>
</tr>
<tr>
<td>OURS</td>
<td>3.32</td>
<td>5.43</td>
</tr>
</tbody>
</table>

68 landmarks

<table>
<thead>
<tr>
<th>TF</th>
<th>FM</th>
<th>300-VW</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESR</td>
<td>3.49</td>
<td>6.74</td>
</tr>
<tr>
<td>SDM</td>
<td>3.80</td>
<td>7.38</td>
</tr>
<tr>
<td>CFAN</td>
<td>3.31</td>
<td>6.47</td>
</tr>
<tr>
<td>TCDCN</td>
<td>3.45</td>
<td>6.92</td>
</tr>
<tr>
<td>PIEFA</td>
<td>3.24</td>
<td>6.07</td>
</tr>
<tr>
<td>OURS</td>
<td>3.17</td>
<td>6.18</td>
</tr>
</tbody>
</table>
Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks

Shengtao XIAO, Jiashi FENG, Junliang XING, Hanjiang LAI, Shuicheng YAN, Ashraf KASSIM

1. Motivation

- Conventional cascaded face shape regression are sensitive to occlusion and large poses
- Necessary to update all points with equal importance?
- Can we ‘annotate’ a given face in a more humanistic way?
  - E.g. knows when and where to annotate

2. Recurrent Attentive-Refinement

(A) Deep feature extraction and robust initialization.

Robust Initialization:

\[ S_0 = \arg \min ||S - S_d||, \text{ s.t. } S \in F \]

\[ S_0 = \arg \min ||S - S_d||_0 + \lambda ||c||_0, \text{ s.t. } S = \sum c_i S_i \]

(B) An attention model in RAR for adaptively selecting key landmark points.

Optimize \( R_a \) to select attention center

\[ R_a = \sum_{t=1}^{\infty} \eta^{t-1} R(S_{t-1}, \hat{S}_t) \]

\[ R(S_{t-1}, \hat{S}_t) = ||\Gamma_t \Delta S_{t-1}||^2_2 - ||\Gamma_t \Delta S_t||^2_2 \]

\[ \Gamma_t = [\gamma_1^t, \gamma_2^t, ..., \gamma_L^t], \text{ with } \gamma_i^t = \kappa \exp\left(\frac{-||S_{t-1}^i - \hat{S}_t||^2_2}{4D^2_t}\right) \]

(C) RAR sequentially refines the landmark estimation.

Refinement Feature Reweighting

\[ \Phi_T(I_t, \hat{S}_t) = [\gamma_1^t \phi_1^t, \gamma_2^t \phi_2^t, ..., \gamma_L^t \phi_L^t] \]

Refinement Loss

\[ L_R^t = ||\Gamma_t (\Delta R S_t - \Delta S_t)||^2_2 \]

Overall Training of RAR:

\[ \sum_{t=1}^{T} \sum_{n=1}^{N} -\eta^{t-1} R_a(S_{t-1,n}, \hat{S}_{t,n}) + L_R^t \]

3. Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>300-W Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Common</td>
</tr>
<tr>
<td>Zhu et.al [2012]</td>
<td>8.22</td>
</tr>
<tr>
<td>RCPR [Burgos, 2013]</td>
<td>6.18</td>
</tr>
<tr>
<td>SDM [Xiong, 2013]</td>
<td>5.57</td>
</tr>
<tr>
<td>LBF [Ren, 2014]</td>
<td>4.95</td>
</tr>
<tr>
<td>LBF Fast [Ren, 2014]</td>
<td>5.38</td>
</tr>
<tr>
<td>CFAN [Zhang, 2014]</td>
<td>5.50</td>
</tr>
<tr>
<td>CFSS [Zhu, 2015]</td>
<td>4.73</td>
</tr>
<tr>
<td>Ours (RAR)</td>
<td>4.12</td>
</tr>
</tbody>
</table>

Select Attention Center such:

\[ C^* = \arg \max_A \text{LSTM}(\Phi_a(I_t, \hat{S}_t); W_a, c) \]
\[ c \in \{1, ..., L\} \]
Ambient Sound Provides Supervision for Visual Learning

Andrew Owens  Jiajun Wu  Josh McDermott  William Freeman  Antonio Torralba

Motivation

Can we learn image representations using ambient sound — instead of manual annotations — as a supervisory signal?

Task: Predict sound from a video frame. To perform this task well, the model should learn to recognize objects and scenes.

Audio representation

We represent audio using sound textures — collections of time-averaged summary statistics [1].

Audio cluster prediction

Images grouped by audio cluster

Clustered audio stats.

Energy

Audio cluster prediction

Images grouped by audio cluster

Clustered audio stats.

Results

The image features that our model learns perform comparably to state-of-the-art unsupervised methods on recognition tasks.

<table>
<thead>
<tr>
<th>Method</th>
<th>VOC Cls (%mAP)</th>
<th>SUN397 (%acc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound (cluster)</td>
<td>36.7 45.8 44.8 44.3</td>
<td>17.3 22.9 20.7 14.9</td>
</tr>
<tr>
<td>Sound (binary)</td>
<td>39.4 46.7 47.4</td>
<td>17.1 22.5 21.3 21.4</td>
</tr>
<tr>
<td>Sound (spectrum only)</td>
<td>35.4 44.0 44.4 44.4</td>
<td>16.6 15.6 18.6</td>
</tr>
<tr>
<td>Texton-CNN</td>
<td>28.9 37.5 35.3 32.5</td>
<td>10.7 15.2 11.4 7.6</td>
</tr>
<tr>
<td>K-means (Krahenbuhl et al.)</td>
<td>27.5 34.8 33.9 32.1</td>
<td>11.6 14.9 12.8 12.4</td>
</tr>
<tr>
<td>Tracking (Wang and Gupta)</td>
<td>33.5 42.2 42.4 42.0</td>
<td>14.1 18.7 16.2 15.4</td>
</tr>
<tr>
<td>Patch pos. (Doersch et al.)</td>
<td>28.8 46.1</td>
<td>- 9.8 22.2</td>
</tr>
<tr>
<td>Egoemotion (Agrawal et al.)</td>
<td>22.7 31.1</td>
<td>- 9.1 11.3</td>
</tr>
<tr>
<td>ImageNet (Krizhevsky et al.)</td>
<td>63.6 65.6 69.6 73.6</td>
<td>29.8 38.0 37.8 37.4</td>
</tr>
<tr>
<td>Places (Zhou et al.)</td>
<td>59.0 63.2 65.3 66.2</td>
<td>39.4 42.1 46.1 48.8</td>
</tr>
</tbody>
</table>

Visualization of the model’s conv5 units (using [2]):

We create a discrete label space by clustering the audio features with k-means, or with an LSH-like binary code.

Grounding of Textual Phrases in Images by Reconstruction

Anna Rohrbach¹, Marcus Rohrbach², Ronghang Hu³, Trevor Darrell², Bernt Schiele¹
¹ Max Planck Institute for Informatics, Saarland Informatics Campus, Germany ² UC Berkeley EECS, Berkeley, CA, United States

Task: Visual Grounding

The two girls in hats in the middle

• Localize arbitrary natural language phrase in an image
• Beyond object detection
• Need to understand the language query

Approach GroundeR: Grounding by Reconstruction

Test time

Predict Attention

A small boy

j = argmax_i \alpha_i

Training time

Attention Loss

\sum \frac{1}{N} \log (P(y, z_i))

Datasets and Experimental setup

• Flickr30k Entities [Plummer 15]
  - 275k bounding boxes & noun phrases
  - Experimental setup
  - 100 object proposals
  - Fast R-CNN [Girshick ICCV’15]

• ReferItGame [Kazemzadeh 14]
  - 99k regions & referring expressions
  - Experimental setup
  - VGG16 + spatial feat [Hu CVPR’16]

Quantitative Results

Flickr30k Entities

<table>
<thead>
<tr>
<th>Approach</th>
<th>Test time</th>
<th>Training time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroundeR (semi-supervised)</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>GroundeR (unsupervised)</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Flickr30k Entities

<table>
<thead>
<tr>
<th>Approach</th>
<th>Test time</th>
<th>Training time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroundeR (semi-supervised)</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>GroundeR (unsupervised)</td>
<td>25%</td>
<td>25%</td>
</tr>
</tbody>
</table>

Multimodal Compact Bilinear (MCB) Pooling: EMNLP’16

Accuracy %: phrases: (OJ(predicted box, ground-truth box) ≥ 0.5)

<table>
<thead>
<tr>
<th>MCB Pooling</th>
<th>Flickr30k Entities</th>
<th>ReferItGame</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroundeR supervised + MCB</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>GroundeR supervised with very little annotations!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Multimodal Compact Bilinear (MCB) Pooling

<table>
<thead>
<tr>
<th>Attack</th>
<th>Flickr30k Entities</th>
<th>ReferItGame</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroundeR supervised + MCB</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>GroundeR supervised with very little annotations!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Qualitative Results

A man in orange pants and brown vest is playing tug-of-war with a dog.

GroundeR supervised (12.5% annot.)

A little brown and white dog emerged from a yellow tabby kitten box onto the lawn.

GroundeR supervised, OSPE [Wang 16]

Is this going to be a feast?

Accuracy in %

<table>
<thead>
<tr>
<th>Approach</th>
<th>Flickr30k Entities</th>
<th>ReferItGame</th>
</tr>
</thead>
<tbody>
<tr>
<td>GroundeR supervised + MCB</td>
<td>40%</td>
<td>30%</td>
</tr>
<tr>
<td>GroundeR supervised with very little annotations!</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GroundeR with two losses: Attention Loss and Reconstruction Loss

• Outperforms the state-of-the-art by a large margin

Conclusions

• Unsupervised grounding (without bounding boxes) is possible!
  - GroundeR with Reconstruction Loss

• Semi-supervised GroundeR works best!
  - GroundeR with two losses: Attention Loss and Reconstruction Loss
    - Efficiently exploits little annotations
    - Outperforms fully supervised GroundeR

For details see [Fukui 16], our EMNLP’16 paper.

References

2. Plummer et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. ICCV’15
3. Donahue et al. Long-term recurrent convolutional networks for visual recognition and description. CVPR’15
5. Hu et al. Natural language object retrieval. CVPR’16
8. SCRC [Hu 16]
9. GroundeR
10. A Semisupervised Challenge
11. Caffe
12. Heatmap
13. LSTM
14. CNN
15. VGG
16. FC
17. LSTM
18. LCNN
19. SCRC
20. [Karpathy 16]
21. Deep Fragments (Karpathy 14)“ going to feast?”
22. [Pham and Page 15] Pham and Page 15, Pham and Page 16, Pham and Page 17, Pham and Page 18. A B C
23. [Plummer 15] Plummer et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. ICCV’15
24. GroundeR supervised + MCB
25. GroundeR supervised
26. Unsupervised
27. Fully supervised
28. Semi-supervised
29. Supervised
30. Unsupervised
31. Fully supervised
32. Semi-supervised
33. Supervised
34. Unsupervised
35. Fully supervised
36. Semi-supervised
37. Supervised
38. Unsupervised
39. Fully supervised
40. Semi-supervised
41. Supervised
42. Unsupervised
43. Fully supervised
44. Semi-supervised
45. Supervised
46. Unsupervised
47. Fully supervised
48. Semi-supervised
49. Supervised
50. Unsupervised
51. Fully supervised
52. Semi-supervised
53. Supervised
54. Unsupervised
55. Fully supervised
56. Semi-supervised
57. Supervised
58. Unsupervised
59. Fully supervised
60. Semi-supervised
61. Supervised
62. Unsupervised
63. Fully supervised
64. Semi-supervised
65. Supervised
66. Unsupervised
67. Fully supervised
68. Semi-supervised
69. Supervised
70. Unsupervised
71. Fully supervised
72. Semi-supervised
73. Supervised
74. Unsupervised
75. Fully supervised
76. Semi-supervised
77. Supervised
78. Unsupervised
79. Fully supervised
80. Semi-supervised
81. Supervised
82. Unsupervised
83. Fully supervised
84. Semi-supervised
85. Supervised
86. Unsupervised
87. Fully supervised
88. Semi-supervised
89. Supervised
90. Unsupervised
91. Fully supervised
92. Semi-supervised
93. Supervised
94. Unsupervised
95. Fully supervised
96. Semi-supervised
97. Supervised
98. Unsupervised
99. Fully supervised
100. Semi-supervised
101. Supervised
102. Unsupervised
103. Fully supervised
104. Semi-supervised
105. Supervised
106. Unsupervised
107. Fully supervised
108. Semi-supervised
109. Supervised
110. Unsupervised
111. Fully supervised
112. Semi-supervised
113. Supervised
114. Unsupervised
115. Fully supervised
116. Semi-supervised
117. Supervised
118. Unsupervised
119. Fully supervised
120. Semi-supervised
121. Supervised
122. Unsupervised
123. Fully supervised
124. Semi-supervised
125. Supervised
126. Unsupervised
127. Fully supervised
128. Semi-supervised
129. Supervised
130. Unsupervised
131. Fully supervised
132. Semi-supervised
133. Supervised
134. Unsupervised
135. Fully supervised
136. Semi-supervised
137. Supervised
138. Unsupervised
139. Fully supervised
140. Semi-supervised
141. Supervised
142. Unsupervised
143. Fully supervised
144. Semi-supervised
145. Supervised
146. Unsupervised
147. Fully supervised
148. Semi-supervised
149. Supervised
150. Unsupervised
151. Fully supervised
152. Semi-supervised
153. Supervised
154. Unsupervised
155. Fully supervised
156. Semi-supervised
157. Supervised
158. Unsupervised
159. Fully supervised
160. Semi-supervised
161. Supervised

Improving Multi-label Learning with Missing Labels by Structured Semantic Correlations
Hao Yang, Joey Tianyi Zhou and Jianfei Cai

Missing Labels Problem
Missing labels problem is common in real world applications, due to
• The number of possible labels/tags could be large.
• There often exists ambiguity among labels

Key Assumptions and Contribution
• Assumption 1: The column vectors in $Y$ lie in the subspace spanned by the column vectors in $X$.
• Assumption 2: Semantically similar images should have similar labels.
1. Propose a simple yet effective semantic vector extraction to project images to semantic space.
2. Exploit structure correlations in the semantic space to boost classification accuracies.
3. Optimizing classification model $W$ instead of label matrix $Y$ to facilitate efficient training

Images with missing labels

Semantic Feature Extraction
Global Semantic Descriptor
Global semantic descriptors are extracted from visual relevant visual concepts from large scale datasets, e.g. ILSVRC, Places.

Local Semantic Descriptor
Local semantic descriptors are generated from labels of visual neighbours.

Experimental Results
Comparison with several baseline methods on four multi-label datasets

Some examples of labels generated using our method

Visual Relationship Detection with Language Priors
Cewu Lu*, Ranjay Krishna*, Michael Bernstein, Li Fei-Fei
Stanford University

Model overview

Motivation
While objects are the core building blocks of an image, it is often the relationships between objects that determine the holistic interpretation.

Challenges
1. Quadratic explosion of label space
   - Detect objects and predicates individually
2. Long-tail distribution of relationships
   - semantic word embeddings

Dataset

Single image output
spatial, comparative, asymmetrical, verb and prepositional relationships

Diverse model predictions

Quantitative results

Zero shot predictions

Visual module
First, we use RCNN to detect all the objects in an image. Next, we take pairs of objects and use our visual module to predict relationships between them:

\[ V(R_i, \Theta) = P_i(O_1) + s_k P_j(O_2) \]

where \( O_1 \) and \( O_2 \) are the subject and object objects. \( \Theta \) is the parameter set of \( z_k, s_k \), the parameters learnt to convert our CNN features to relationship likelihoods.

Language module
We score every relationship using their word2vec features:

\[ f(R, W) = w_k^T [word2vec(t_1), word2vec(t_2)] + b_k \]

where \( w_k \) and \( b_k \) are the subject and object words. \( W \) is the set of \( \{w_1, b_1, \ldots, w_K, b_K\} \), where each row presents one of our \( K \) predicates.

We train \( W \) by ensuring that similar relationships are projected closer together:

\[ K(W) = \text{var}(\{ f(R, W) - f(R', W) \}^2 \forall R, R' ) \]

where \( d(R, R') \) is the sum of the cosine distances (in word2vec space) between the two objects and the predicates of the relationships. var() is the variance function.

Finally, to ensure that \( f(R) \) outputs the likelihood of a relationship, we learn which relationships are more probable using a ranking loss function:

\[ L(W) = \sum_{R, R'} \max \{ f(R, W) - f(R, W) + 1, 0 \} \]

So our final objective function becomes:

\[ \min_{\Theta, W} \{ C(\Theta, W) + \lambda_1 L(W) + \lambda_2 K(W) \} \]

where

\[ C(\Theta, W) = \sum_{\{O_1, O_2\}, R} \max \{ 1 + V(R, \Theta)(O_1, O_2) f(R, W), 0 \} \]

\[ + \sum_{\{O_1, O_2\}, R} \max \{ 1 + V(R, \Theta)(O_1, O_2) f(R', W), 0 \} \]

\[ \forall R' \neq R \]
An Efficient Fusion Move Algorithm for the Minimum Cost Lifted Multicut Problem

Thorsten Beier* Björn Andres† Ulrich Köthe* Fred A. Hamprecht*  
† Computer Vision and Multimodal Computing, Max Planck Institute for Informatics

Introduction

The Minimum Lifted Multicut Problem:

● An Optimization problem whose feasible solutions are decompositions of a graph
● Objective function can penalize or reward all decompositions for which any given pair of nodes are in distinct components
● We propose a fusion move algorithm which outperforms existing algorithms
● We use this objective function for image segmentation and obtain a new state of the art for a problem in biological image analysis

Multicut Objective [4, 3, 6]

The minimum multicut finds the clustering which minimizes the sum of weights between clusters. Given a graph \( G = (V, E) \) with edge weights \( c : E \to \mathbb{R} \) this can be formulated as an ILP:

\[
\begin{align*}
\min \quad & \sum_{e \in E} c_e x_e \\
\text{subject to} \quad & \forall Y \in \text{cycles}(G) \forall e \in Y : \quad x_e \leq \sum_{e' \in Y} x_{e'}
\end{align*}
\]

\( x_e \) must be cut \( \iff \) \( e \) is part of a cycle.

Fusion Move Algorithm for Lifted Multicuts

We show how to implement a fusion moves alg.[8, 5] for the lifted multicut objective:

\( \text{Solution } A \) + \( \text{Solution } B \) = \( \text{Optimal lifted multicut} \)

Must link constraints for edges which are uncut in \( A \) and \( B \)

Figure: Left Figures: Two solutions of the lifted multicut problem \( A \) and \( B \) can be fused into a single solution by contracting all edges which are uncut in \( A \) and \( B \) (lifted edges in \( E^t \setminus E \) are depicted in green). We can optimize the lifted multicut problem on the smaller contracted graph to get the fused solution. Right Figures: In a fusion move algorithm, proposal generation (PG) and fusion moves (FM) can be combined in different ways. We implement and study serial fusion moves (top) and parallel fusion moves (bottom)

Source Code available:  
https://github.com/DerThorsten/nifty

Results: ISBI Challenge 2012

Segmentation of Neuronal Processes in EM Images:

● Learn membrane probability map with CNN
● Generate superpixels
● Learn cut-probabilities for local edges (A-B, C-D) and lifted edges (E-F, G-H)
● Optimize with the proposed algorithm

Complete Pipeline / Source Code available:  
https://github.com/DerThorsten/lifted_fusion_moves_eccv_2016

Results: Benchmarks

Table: Left: Performance on the large and hard instances of the minimum cost lifted multicut problem of [7]. Right: performance on instances of a generalization of [1]

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Objective</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1,2,4,8 threads)</td>
</tr>
<tr>
<td>our Alg.</td>
<td>-627447</td>
<td>10.8</td>
</tr>
<tr>
<td>KL [8]</td>
<td>-627455</td>
<td>121</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Objective</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(1/2/4/8 Threads)</td>
</tr>
<tr>
<td>our Alg.</td>
<td>-2.29e+07</td>
<td>14.8</td>
</tr>
<tr>
<td>GAEC + KLj</td>
<td>-2.27e+07</td>
<td>29.3</td>
</tr>
</tbody>
</table>

References

Introduction

- High dimensional data often lie in a set of low-dimensional subspaces in many practical scenarios.
- Subspace clustering: partition the data such that data belonging to the same subspace are identified as one cluster.
- Among various subspace clustering algorithms, the ones that employ sparsity prior, such as Sparse Subspace Clustering (SSC), are effective in subspace clustering under certain assumptions.
- Typical algorithms build similarity matrix in accordance with the subspace-sparse representation, then apply spectral clustering on this similarity matrix to obtain clustering result.
- We present $\ell^0$-Sparse Subspace Clustering ($\ell^0$-SSC) with theoretical advantages and compelling empirical results compared to SSC and other competing subspace clustering methods.

Contributions

Key element: subspace-sparse representation or the representation that satisfies subspace detection property, which specifies data in the same subspace for each data point.

Main contributions of $\ell^0$-SSC:
- Almost surely equivalence between $\ell^0$-sparsity and the subspace detection property, under the mildest assumption to the best of our knowledge.
- Although the optimization of $\ell^0$-SSC is NP-hard, we present an efficient proximal algorithm with theoretical guarantee. More concretely, under certain assumptions on the sparse eigenvalues of the data, the proximal algorithm guarantees convergence to the critical point of the objective, and the obtained sub-optimal solution is close to the globally optimal solution.

Formulation

Theorem 1 ($\ell^0$-sparsity $\Rightarrow$ subspace detection property) Suppose the data $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ lie in a union of $K$ distinct subspaces $\{S_k\}_{k=1}^K$ of dimensions $\{d_k\}_{k=1}^K$. Let $X^{(k)} = \mathbb{R}^{d \times n_k}$ denote the data that belong to subspace $S_k$, and $\sum_{k=1}^K n_k = n$. When $n_k \geq d_k + 1$, if the data belonging to each subspace are generated i.i.d. from arbitrary unknown continuous distribution supported on that subspace, then with probability 1, the optimal solution to (1) satisfies the subspace detection property, i.e. nonzero elements of $\alpha^*$ corresponds to the data that lie in the same subspace as $x_i$.

Illustration of a inter-subspace hyperplane. The hyperplane spanned by $x_i \in S_1$ and $x_j \in S_2$ is a inter-subspace hyperplane, and the intersection of this inter-subspace hyperplane and $S_1$ is the dashed line $x_i \cdot O \cdot A$.

Experimental Results

Clustering result on several face data sets, with comparison to several competing methods. KM: K-means, SC: Spectral Clustering, SMCE: Sparse Manifold Clustering and Embedding, SSC-OMP: using Orthogonal Matching Pursuit (OMP) to solve (1).
Normalized Cut meets MRF

Meng Tang¹ Dmitrii Marin¹
¹University of Western Ontario, Canada

Ismail Ben Ayed² Yuri Boykov¹
²École de Technologie Supérieure, Canada

Summary of contribution
- new unary kernel and spectral bounds for NC
- can combine NC with any MRF constraints
- can combine MRF with balanced clustering
- MRF with features of any dimension (RGBD, RGBM, RGBXYM, deep,...)

Experiments: NC helps MRF
RGBD segmentation
- poor scalability to high-dim

Experiments: MRF helps NC
Potts model improves edge alignment

Energy Formulation
Normalized Cut (NC)
\[ E(S) = -\sum_k \frac{\text{assoc}(S_k, S_k)}{\text{assoc}(S_k, \Omega)} + \gamma \sum_{c \in \mathcal{C}} E_c(S_c) \]
MRF
\[ \sum_{p \in \Omega} U_p(S_p) + \gamma \sum_{c \in \mathcal{F}} E_c(S_c) \]

Why MRF for NC
- weak edge alignment
- semi-supervision is challenging
- how to incorporate group priors?
- how many clusters?

Why NC for MRF
- high dimension feature: RGBXYM (location XY, motion M)
- can combine MRF with balanced clustering
- MRF with features of any dimension (RGBD, RGBM, RGBXYM, deep,...)

Why NC for MRF (cont.)
- low-rank (m) approximation
- spectral bound for NC (approx.)
- spectral bound accuracy

Optimization: Kernel Cut and spectral Cut
General optimization approach: bound optimization
\[ A_f(S) := \sum_{p \in \Omega} U_p(S_p) + \gamma \sum_{c \in \mathcal{F}} E_c(S_c) \]
to optimize bound \( A_f(S) \) we use move-making graph cuts

Kernel bound for NC (exact)
- Normalized Cut is concave:
- 1st order Taylor expansion

Spectral bound for NC (approx.)
Step 1. low-rank (m) approximation
\[ \min \| A - \hat{A} \|_F \]
Step 2. low-dim. isometric Euclidean embedding
\[ \hat{A}_{m} = \sqrt{\lambda_m} \hat{V}' \hat{U} \]

Spectral bound accuracy
- (depending on m)

Experiments: NC helps MRF
RGBD segmentation
- poor scalability to high-dim

Experiments: MRF helps NC
Potts model improves edge alignment

Energy Formulation
Normalized Cut (NC)
\[ E(S) = -\sum_k \frac{\text{assoc}(S_k, S_k)}{\text{assoc}(S_k, \Omega)} + \gamma \sum_{c \in \mathcal{C}} E_c(S_c) \]
MRF
\[ \sum_{p \in \Omega} U_p(S_p) + \gamma \sum_{c \in \mathcal{F}} E_c(S_c) \]

Why MRF for NC
- weak edge alignment
- semi-supervision is challenging
- how to incorporate group priors?
- how many clusters?

Why NC for MRF
- high dimension feature: RGBXYM (location XY, motion M)
- can combine MRF with balanced clustering
- MRF with features of any dimension (RGBD, RGBM, RGBXYM, deep,...)

Why NC for MRF (cont.)
- low-rank (m) approximation
- spectral bound for NC (approx.)
- spectral bound accuracy

Optimization: Kernel Cut and spectral Cut
General optimization approach: bound optimization
\[ A_f(S) := \sum_{p \in \Omega} U_p(S_p) + \gamma \sum_{c \in \mathcal{F}} E_c(S_c) \]
to optimize bound \( A_f(S) \) we use move-making graph cuts

Kernel bound for NC (exact)
- Normalized Cut is concave:
- 1st order Taylor expansion

Spectral bound for NC (approx.)
Step 1. low-rank (m) approximation
\[ \min \| A - \hat{A} \|_F \]
Step 2. low-dim. isometric Euclidean embedding
\[ \hat{A}_{m} = \sqrt{\lambda_m} \hat{V}' \hat{U} \]

Spectral bound accuracy
- (depending on m)

Experiments: NC helps MRF
RGBD segmentation
- poor scalability to high-dim

Experiments: MRF helps NC
Potts model improves edge alignment
Phase-based Modification Transfer for Video

Simone Meyer, Alexander Sorkine-Hornung, Markus Gross
ETH Zurich / Disney Research

Motivation
Goal: Propagating modifications of one frame to subsequent video frames.
- Existing methods slow due to computing pixel correspondences and global optimization.
- Method should be able to handle different kind of modifications without any additional information.
- Focus on efficiency, enabling its application to high resolution and high frame rate content.
- Investigating, whether modification transfer without explicit correspondences is possible.

Our Approach
Our method uses the assumption that small motion can be represented as phase-shift.
[Portilla and Simoncelli 2000]
- Our approach only uses per pixel modifications.
- Its implementation is fast and suitable for parallelization.

Our Contributions
Basic approach
Correction of Phase Differences
Correction of Amplitudes

Algorithm
Problem: Changed frequency content
Step 0: Decomposition of images and computation of phase values
\[ R_{m,n}(x,y) = (f + R_{m,n})(x,y) = \left( C_{m,n}(x,y) + I_{m,n}(x,y) \right) = A_{m,n}(x,y) \cdot e^{i \phi_{m,n}(x,y)} \]

Step 1: Detection and correction of missing phase information
- Detection of relevant motion
\[ \phi_1(A_{m,n}(x,y), A_{m,n}(y)) = \max(A_{m,n}(x,y), A_{m,n}(y)) - \phi_1 \]
- Detection of significant modifications
\[ \phi_2(A_{m,n}(x,y), A_{m,n}(y)) = \left( A_{m,n}(x,y) - A_{m,n}(y) \right) - \phi_2 \]
- Correction of phase differences
\[ \dot{\phi}(x,y) = \left\{ \begin{array}{ll} \dot{\phi}_{m,n} - \lambda(t) \dot{\phi}_m & \text{if } (v_1 > \tau_2) \land (v_2 < \tau_2) \\ \dot{\phi}_{m,n} - \dot{\phi}_m & \text{otherwise} \end{array} \right\} \]

Step 2: Correction of amplitudes
- Magnification factor
\[ \eta(A_1(x,y)) = \left\{ \begin{array}{ll} \frac{\max(A_{m,n})}{\max(A_{m,n})} & \text{if } \eta \end{array} \right\} \]

Overview
- Pyramid decomposition
- Our algorithms

Results
Basic approach
Our result
Input images
Optical flow [Brox et al.]
Our result

Input images
Modified input, t=0
Our result, t=5

Conclusions
Our method describes a phase-based approach for modification transfer, only using per-pixel phase modifications and is therefore suitable to process high resolution and high frame rate content.
**Problem Statement**
Given a grayscale image, predict the color.

**Qualitative Comparisons**

<table>
<thead>
<tr>
<th>Success Cases</th>
<th>Classification</th>
<th>L2 Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grayscale</td>
<td>Ground Truth</td>
<td>Rebal</td>
</tr>
</tbody>
</table>

**Failure Cases**

<table>
<thead>
<tr>
<th>Grayscale</th>
<th>Classification</th>
<th>L2 Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>Rebal</td>
<td>Ground Truth</td>
</tr>
</tbody>
</table>

**Our Contributions**

1. **Graphics Task of Colorization**
   - Achieve state-of-the-art by training on 1M ImageNet photos
   - Design an appropriate objective function that handles the multimodal uncertainty and captures a wide diversity
   - Introduce a novel framework for testing colorization algorithms, potentially applicable to other image synthesis tasks

2. **Colorization as Representation Learning**
   - Introduce colorization task as instance of cross-channel encoding
   - Evaluate colorization for representation learning, demonstrate competitive performance in self-supervision framework

**Inherent Ambiguity**
Multiple plausible colorizations may exist.

- L2 loss is inadequate for this problem

**Our Loss Function**

Grayscale image to color distribution:
- Quantize ab space into grid size 10, keep 313 bins in gamut
- Cross entropy loss

\[
L(Z, Z') = -\frac{1}{n} \sum_{i=1}^{n} \left( y_i \log p_i + (1 - y_i) \log (1 - p_i) \right)
\]

- Class rebalancing to encourage learning of rare colors

\[

\text{Label} = \frac{1}{(1 + \beta)} \cdot \sum_{i} p_i \beta_i = 1
\]

- Weighted empirical distribution combine with uniform log, probability

**Network Architecture**

Fully convolutional architecture, VGG-style

**Quantitative Comparisons**

Use 3 metrics of evaluation:
1. Per-pixel accuracy (AUC-CMF)
2. Commonly used metric for colorization
3. Does not evaluate plausibility, or joint interaction between pixels
4. Semantic interpretability (VGG)
5. Perceptual realism (AMT)

**Colorization Results on ImageNet**

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC-CMF</th>
<th>VGG 100-1</th>
<th>AMT-Labeled-R</th>
<th>AMT-Labeled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground Truth</td>
<td>99.0</td>
<td>98.0</td>
<td>98.0</td>
<td>98.0</td>
</tr>
<tr>
<td>Gray</td>
<td>98.0</td>
<td>97.0</td>
<td>97.0</td>
<td>97.0</td>
</tr>
<tr>
<td>Lemaire et al. [20]</td>
<td>95.0</td>
<td>94.0</td>
<td>94.0</td>
<td>94.0</td>
</tr>
</tbody>
</table>

**Task Generalization**

- Fine-tune AlexNet features for PASCAL tasks

**Per-Pixel Color Distribution to Single Point Estimate**

Mean: Spatially coherent but disattenuated
- Mode is vibrant but can have artifacts
- Interpolate between mean and mode with annealed-mean

\[
H(\mathbf{Z}^{(a)}) = \mathbf{f}(\gamma \log \mathbf{Z}^{(a)}), \quad \gamma = \frac{\exp(a/T)}{\log \exp(a/T)}
\]

**Semantic Interpretability of Results (VGG Classification)**

**Representation Learning via Cross-Channel Encoding**

Supervised training: X₀ → CNN → Label y

Unsupervised/Self-supervised training: X₀ → CNN → X₁

**Dataset & Task Generalization**

How does network generalize to unseen data?
- Fine-tune AlexNet features for PASCAL tasks
Measuring Depth and Velocity with Defocus and Differential Motion

**Motivation**  
Low power (mW) depth sensing

![Image](rubenstein_et_al_14.png)  
![Image](ma_et_al_13.png)

~200 mW  
~20 mW

---

**Idea**  
Combine motion and defocus blur

Wide aperture (Thin-lens Model)

![Image](lost_and_taken.png)

Texture plane

Pinhole

Optical Flow

\[
\begin{bmatrix}
\{\dot{x}, \dot{y}\}
\end{bmatrix} \begin{bmatrix}
P_x & P_y & P_t
\end{bmatrix} = 0
\]

Derivation

Gaussian blur reveals depth

\[
l_i + x l_x + y l_y = R(Z, \dot{x}, \dot{y}, P; \kappa, Z, \hat{r}, \hat{m}) = \frac{Z}{Z - \hat{r}} \left( \frac{1}{Z} - \frac{r}{\sigma(Z)} \right) \left( 2k \left( \frac{r}{\sigma(Z)} \right) + r \kappa \left( \frac{r}{\sigma(Z)} \right) \right) \ast P
\]

\[
F(Z, \dot{x}, \dot{y}, \hat{r}, \hat{m}) = m \ast k \ast r
\]

Texture independence

Fourier transform

Solve differential equation

Compact operator

Nonnegative transmittance

Inverse Fourier Transform

\[
0 = [I_x, I_y, (x I_x + y I_y) \ast \hat{m}] \ast \hat{u} = \hat{u} = 0
\]

Proof of Concept

Experimental results

---

**Contribution**  
Optical Flow

Focal Flow

Image Motion

Depth & 3D Velocity

\[
A_{4 \times 4} \hat{u} = \hat{b}_{4 \times 1}
\]

\[
Z_{i, j, k, \hat{r}, \hat{m}}
\]

\[
\frac{\dot{X}}{\dot{Y}} = \frac{Z_{i, j, k, \hat{r}, \hat{m}}}{Z_{i, j, k, \hat{r}, \hat{m}}}
\]

\[
\frac{\dot{X}}{\dot{Y}} = \frac{Z_{i, j, k, \hat{r}, \hat{m}}}{Z_{i, j, k, \hat{r}, \hat{m}}}
\]

---

**Derivation**

Gaussian blur reveals depth

Filter \( \kappa(r) \)

\[
l_i + x l_x + y l_y = R(Z, \dot{x}, \dot{y}, P; \kappa, Z, \hat{r}, \hat{m}) = \frac{Z}{Z - \hat{r}} \left( \frac{1}{Z} - \frac{r}{\sigma(Z)} \right) \left( 2k \left( \frac{r}{\sigma(Z)} \right) + r \kappa \left( \frac{r}{\sigma(Z)} \right) \right) \ast P
\]

\[
R(Z, \dot{x}, \dot{y}, P; \kappa, Z, \hat{r}, \hat{m}) = v(Z, \dot{x}, \dot{y}, P; \kappa, Z, \hat{r}, \hat{m}) \ast P
\]

Texture independence

Fourier transform

Solve differential equation

Compact operator

Nonnegative transmittance

Inverse Fourier Transform

\[
0 = [I_x, I_y, (x I_x + y I_y) \ast \hat{m}] \ast \hat{u} = 0
\]

---

**Experimental results**

Code, equipment, results: https://vision.seas.harvard.edu/focalflow
Top-down Neural Attention by Excitation Backprop

Jianming Zhang¹, Zhe Lin¹, Jonathan Brandt¹, Xiaohui Shen¹, Stan Sclaroff²
¹Adobe Research  ²Boston University

Motivation: Visualizing CNNs’ Top-Down Attention

- Problem
  - Bottom-up Inference

- Example
  - Top-down Attention Maps of a 18K-Tag Classifier Using Our Method

Method

- Beyond Selective Tuning: Probabilistic Winner-Take-All
  - Winner-Take-All [1]
  - Marginal Winning Probability (MWP)

- Excitation Backprop
  - Assumptions:
    1) The responses of the activation neurons are non-negative.
    2) An activation neuron is tuned to detect certain visual features. Its response is positively correlated to its confidence of the detection.

Experiments

- The Pointing Game
  - Task:
    - Given an image and an object category, point to the targets.
  - Metric:
    - Mean pointing accuracy across categories.
    - Pointing anywhere on the targets is fine.

- Top-down Attention from a 18K-Tag Classifier
  - An image tag classifier for 18K tags
    - Trained on 6M Stock images with user tags
    - Cross entropy multi-label loss
  - Phrase Localization
    - Evaluate on the Flickr 30K entity dataset [2]

Reference

Learning Recursive Filters for Low-Level Vision via a Hybrid Neural Network

Sifei Liu, Jinshan Pan, Ming-Hsuan Yang
University of California at Merced
http://www.sifeiliu.net/project

**Introduction**

**Motivation:** Numerous low-level vision problems (e.g., edge-preserving filtering and denoising) can be regarded as recursive image filtering via a hybrid neural network.

**Goal:** Learning recursive filters via a deep CNN
- General framework for low-level vision tasks
- A data-driven model which does not consider the original models and the complexity of implementation details

**Convolutional Filter v.s. Recursive Filter**

**Convolutional Filter**

\[ y[k] = \sum_{i=0}^{N} a_i x[i] \]

- Easy to design
- Large number of parameters
- Many groups of filters

**Recursive Filter**

\[ y[k] = a_0 x[k] + p y[k - 1] \]

- Small number of parameters
- Difficult to design

**Recursive is Linear Recurrent**

- Temporal domain
- Z domain

A general recursive filter

\[ y[k] = \sum_{i=0}^{N} a_i x[i] + p y[k - 1] \]

A recursive unit

\[ y[k] = p y[k] + p y[k - 1] \]

**Spatially Variant Linear RNN**

**Proposed Linear RNN**

\[ y[k] = g x[k] + py[k - 1] \]

- Normalized filter \( g = 1 - p \)
- Spatially variant \( y[k] = (1 - p) x[k] + py[k - 1] \)
- Form of back propagation \( \sigma[k] = \theta(y[k - 1] - x[k]) \)

**Joint Training**

**CNN: Weight Map Learning**

**LRNN: Filtering**

**Weight Map Visualization via 1st Order LRNN**

**Learning Recursive Filters**

**Edge-Preserving Filtering**

**Image Denoising**

**Inpainting**

**Colorization**

**Re-colorization**

**Run Time**

**Conclusions**

- A novel hybrid neural network for low-level vision tasks
- Faster speed and smaller model size

---

**Figures:**
- Smoothing
- Denoising
- Inpainting
- Color interpolation
- Convolutional Filter v.s. Recursive Filter
- Convolutional Filter
- Recursive Filter
- Recursive is Linear Recurrent
- Spatially Variant Linear RNN
- Joint Training
- CNN: Weight Map Learning
- LRNN: Filtering
- Weight Map Visualization via 1st Order LRNN
- Learning Recursive Filters
- Edge-Preserving Filtering
- Image Denoising
- Inpainting
- Colorization
- Re-colorization
- Run Time
- Conclusions
Learning Representations for Automatic Colorization

Gustav Larsson  
University of Chicago

Michael Maire  
TTI Chicago

Greg Shakhnarovich  
TTI Chicago

Overview

Problem statement:
- Input: Grayscale image
- Output: Plausible and pleasing color rendition

Background
- Three methods in the literature:
  - Scribble based: User colorizes some pixels, the algorithm fills in the rest
  - Transfer based: Color is transferred from a reference image
  - Fully automatic: Parametric model that predicts color directly
- Recent interest in fully automatic colorization:
  - Deshpande et al. (ICCV 2015), Cheng et al. (ICCV 2015), Izitka (SIGGRAPH 2016), Zhang et al. (ECCV 2016)

Our work
- Design principles:
  - Semantic cues: Leverage ImageNet-based classifier
  - Low-level/high-level cues: Zoom-out/hypercolumn architecture
  - Colorization not unique: Predict color histograms
- Fully automatic: Takes less than 1 second per image.
- High-quality results. State-of-the-art on all quantitative measures.

Experimental Results

Dense Hypercolumns
- Low-level layers are upsampled
- High memory footprint
- Sparse
- Direct bilinear sampling
- Low memory footprint

Examples

Automatically colorized photos

Failure modes
- Colorize.ttic.edu
- gustavla/autocolorize

pip install autocolorize

autocolorize graysc.png -o color.png

Network Architecture

Sparse Training
- Dense
- Low-level layers are upsampled
- High memory footprint
- Sparse
- Direct bilinear sampling
- Low memory footprint

Table: Metrics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>RMSE</th>
<th>PSNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>0.285</td>
<td>25.05</td>
</tr>
<tr>
<td>SUN-6</td>
<td>0.211</td>
<td>26.05</td>
</tr>
<tr>
<td>SUN-6 (GT Hist)</td>
<td>0.178</td>
<td></td>
</tr>
</tbody>
</table>

Code available for Caffe/TensorFlow

Failure modes

Sampling multiple colorizations using the rich histogram representation's color uncertainty

Failure modes

Failure modes

Examples

Automatically colorized photos

Old B&W photographs automatically colorized

Training

We employ a fully convolutional network and consider several training formulations:
- Color space
- Loss function
- Spatial sampling
- Histogram predictions (K bins)

We chose hue/chroma with histogram predictions at sparse locations.

Table: Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>1.3M</td>
<td>1000</td>
</tr>
<tr>
<td>cval1k</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>SUN Database</td>
<td>1.3M</td>
<td>1000</td>
</tr>
<tr>
<td>SUN-A</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>SUN-6</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

Input: Grayscale Image  
Output: Color Image

Network Architecture

VGG-16-Gray

Hypercolumns

Hue

Ground-truth

Dense Hypercolumns

Sparse Hypercolumns

• Dense

• Allows larger images / more images per batch

• Widely applicable to image-to-image tasks

• Code available for Caffe/TensorFlow
Real-Time 3D Reconstruction and 6-DoF Tracking with an Event Camera
Hanme Kim, Stefan Leutenegger, and Andrew J. Davison
Imperial College London, UK

Motivation

- Visual SLAM is a key enabler for many applications such as AR/VR, drone, autonomous vehicle, and robotics.
- Current visual SLAM products still suffer from motion blur, low dynamic range, high processing time, and high power consumption.
- Event cameras have the potential to overcome all of the issues by reporting only scene changes with high temporal resolution and high dynamic range properties.
- Can we use event cameras for SLAM problems?
- So far, this has only been successfully achieved in the reduced case of restricted motion and 2D reconstruction. (e.g. Kim et al., Simultaneous Mosaicing and Tracking with an Event Camera, BMVC 2014)

Method Overview

- Takes inputs in the form of a stream of events and processes event-by-event.
- Consists of 3 decoupled probabilistic filters and parallel log intensity reconstruction.
- Virtual keyframe:
  - projective reference frame
  - consists of gradient, intensity and depth
  - all probabilistic filters are estimated relative to this
  - uses a higher resolution than for the low resolution sensor
  - textured semi-dense 3D point cloud can be generated from this

Technical Details

- Assume that the current estimates from other components are accurate enough to lock for estimating the other.
- EKF based 6-DoF event camera pose estimation: constant position motion model
  \[ h_e(x(t)|x_{t-1}) = \frac{1}{2} t \]
- Pixel-wise EKF based log intensity gradient estimation:
  \[ h_{I} = \frac{1}{2} t \]
- Pixel-wise EKF based inverse depth estimation:
  \[ h_{I} = \frac{1}{2} t \]
- Parallel log intensity reconstruction:
  \[ \min \{ \sum_{i} (I(x) - \nabla I(x)) (\nabla I(x))^{T} + \lambda (|\nabla I(x)|)^{2} \} \]

Results (please also refer to our supplementary video at https://youtu.be/yHLyhdM5w7w)

Single Keyframe Estimation

Multiple Keyframes

High Speed Tracking

High Dynamic Range

Video Rendering
**Single Image 3D Interpreter Network**

Problem: 3D structure and pose estimation from a single RGB image

Challenge: 3D annotations are hard to obtain

Solution: Use synthetic 3D object models for training

Challenge: Hard to render realistic images with synthetic 3D data

Solution: Use heatmaps of 2D keypoints as intermediate representations

Challenge: Errors propagate in a two-stage model

Solution: Add a 3D to 2D projection layer for end-to-end finetuning

---

**Overview**

**Problem:** 3D structure and pose estimation from a single RGB image

**Challenge:** 3D annotations are hard to obtain

**Solution:** Use synthetic 3D object models for training

**Challenge:** Hard to render realistic images with synthetic 3D data

**Solution:** Use heatmaps of 2D keypoints as intermediate representations

**Challenge:** Errors propagate in a two-stage model

**Solution:** Add a 3D to 2D projection layer for end-to-end finetuning

---

**3D INterpreter Network (3D-INN)**

**2D Annotated Images**

**Initial Heatmaps**

**Revised Heatmaps**

**3D Synthetic Data**

**2D Coordinates**

**Pipeline**

- Input image
- Keypoint heatmaps
- Reconstructed 3D Structure
- 3D Parameters
- 3D Skeletons

**Network Connection**

- **Three-stage training:**
  1) Image to 2D keypoint
  2) 2D keypoint to 3D skeleton
  3) End-to-end fine-tuning

---

**Evaluation**

**3D Structure and Pose Estimation**

**2D Keypoint Estimation**

**Visualization**

**Retrieval**

- By 3D-INN structure
- By 3D-INN viewpoint

---

**Object Graph**

**Contributions**

- **3D-INN:** a generative model estimating 3D structure/pose from a single image
- Connecting 2D annotations and synthetic 3D objects via heatmaps of keypoints
- Enabling end-to-end training w/o 3D labels through a 3D-to-2D projection layer
Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?

Why Use a 1D Sensor for 3D Scanning?
Shape acquisition and registration for 3D endoscope based on grid pattern projection

Ryo Furukawa†, Hiroki Morinaga††, Yoji Sanomura†††, Shinji Tanaka††††, Shigeto Yoshida†††† and Hiroshi Kawasaki†††
†Hiroshima City University, ††Kagoshima University, †††Hirosima University, †††† Hiroshima General Hospital of West JR

- Development of 3D endoscope system based on active stereo
- Sharp and deep depth-of-field pattern projector using DOE
- Pattern features that are robust to blurring by subsurface scattering of bio-tissues
- Rigid and non-rigid registration for sparse 3D shapes

Introduction

Needs for measuring the size of the tumor in endoscopic diagnosis

Development of 3D endoscope system based on active stereo
- Successfully measured ex vivo tumor samples with normal endoscope with an attachment of a micro pattern projector [1].

Problems
- Bio-tissue surfaces have strong sub-surface scattering that causes blurring of projected patterns and reduction of light energy
- Using largely-spaced grid pattern result in coarse point clouds

System configuration

- DOE: Diffractive Optical Element is an optical element that generates patterns by diffraction of laser light
- Local sub-graph patterns (LSPs) are used as matching template to ignore minor topological errors.
- Detected grid graph includes erroneous edge detections or erroneous gap codes.
- After all LSPs are scanned for those criteria fulfills a matching condition, correspondences are voted for each nodes of LSP.

Experiments and evaluation

- For aligning grid-structured point sets, naïve ICP often fails, because parallel lines from different frames attract each other and erroneously forms bundled set of lines (Result A).
- By using closest point pairs between vertical and horizontal point sets from different frames, better alignment results are obtained (Result B).

Aligning grid-structured point sets to obtain dense 3D shapes

- Preprocess of endoscope image
- Detection of vertical lines (BP method)
- Detection of candidate horizontal line segments
- Detection of horizontal line segments
- Extraction grid structure and codes of line-segment gaps
- Matching by local sub-graph

Subgraph-based matching

- Detected grid graph includes erroneous edge detections or erroneous gap codes.
- The detected grid graph G and the grid graph of the original pattern P is compared in subgraph topology of the LSP.
- Comparison criteria are node-wise epipolar constraints, and node-wise gap-code error rates in LSP. If those criteria fulfills a matching condition, correspondences are voted for each nodes of LSP.
- Multiple LSPs are provided to account for various topological errors.
- After all LSPs are scanned for G and P, correspondences are decided from vote results.

- Features of the proposed pattern
- Pattern codes represented by area size of box
- New frequency pattern modulation

- Light energy loss < 5%
- Laser light is generated by DOE (Diffractive Optical Element) and is an optical element that generates patterns by diffraction of laser light
- Deep depth of field (same as laser beam)
- Light energy loss < 5%

- Endoscope head and Instrument channel
- Laser light and Optical fiber
- DOE pattern projector

- Wave pattern [3]
- Microstructure → easily blurred out by subsurface scattering

- Grid ICP (front, top, and side views)
- Non-interpolated, aligned 3D shapes by Interpolated 3D shapes
- (detailed shapes are lost)

- Common ICP
- Interpolated 3D shapes
- (detailed shapes are lost)

- Captured image
- Detected codes
- Interpolated 3D shapes
- (detailed shapes are lost)
Spot On: Action Localization from Pointly-Supervised Proposals

Pascal Mettes
University of Amsterdam

Jan C. van Gemert
Delft University of Technology

Cees G.M. Snoek
University of Amsterdam

Introduction

Discover what actions occur when and where in videos.

Current bottleneck in action localization:
Annotating bounding boxes for each frame of each action is expensive. Therefore, current datasets are sparse and with few videos.

Our hypothesis:
Training on bounding box annotations is not required. Training on unsupervised proposals with fast point annotations is as effective.

Method

Method overview

Start from points annotated on the frames containing the action. Train classifier on best action proposals, guided by point annotations.

Mining the best proposals

We start from Multiple Instance Learning and incorporate information from points. We compute an affinity between each proposal and the point annotations. We iteratively select best proposals (using affinities as prior) and retrain classifiers.

Formal objective:
\[
\min_{w,b,\xi} \frac{1}{2} ||w||^2 + \lambda \sum_i \xi_i \quad \text{s.t.} \quad \forall_i : y_i \cdot (w \cdot \arg\max_{x \in X} P(x|w,b,A_i,C_i,V_i) + b) \geq 1 - \xi_i, \\
\forall_i : \xi_i \geq 0
\]

Proposal selection function:
\[
P \left( \frac{x \cdot w}{b}, A_i, C, V_i \right) \propto (w \cdot z) + b \cdot O \left( A_i, C, V_i \right)
\]

Select the proposal with the best combined classifier score and affinity score.

Proposal and Point Affinity

Novel overlap measure between a proposal (A) and points (C) in video (V).

\[
O(A, C, V) = M(A, C) - S(A, V)
\]

Match score:
Notion: point annotations should be near the center of the proposal boxes. The match score is the inverted distance between points and proposal centers. The distance is divided by the distance between the proposal center and edge.

Size regularization:
The relative size of the proposal (A), compared to the whole video (V).

\[
S(A, V) = \left( \frac{\sum_{i=1}^N ||A_i||}{||V||} \right)^2
\]

Results (UCF Sports, UCF 101 in paper)

1: Training on proposals vs. ground truth boxes

Best possible proposal performs similar to full ground truth boxes.

Scores maintained with point-supervised proposals.

Results holds across overlaps and datasets.

2: Lowering the annotation frame rate

Points are 15 to 50 times faster to annotate than bounding boxes.

Scores are maintained when annotating 10% of the frames.

4: Comparison to state-of-the-art

Method

Supervision

AUC

Lan et al. ICCV'11
box
0.380

Tian et al. CVPR'13
box
0.420

Jain et al. CVPR'14
box
0.489

van Gemert et al. BMVC'15
box
0.546

Soomro et al. ICCV'15
box
0.550

Gkioxari et al. CVPR'15
box
0.559

Weinzaepfel et al. ICCV'15
box
0.559

Jain et al. ICCV'15
zero-shot
0.232

Cinbis et al. CVPR'14
video label
0.278

This work
point
0.545

Both use same action proposals and features.

Points are competitive to boxes, improve over other weak supervisions.

3: Introducing Hollywood2Tubes

New dataset to demonstrate how easy action annotation becomes.
Contains several actions and instances new to action localization.

tinyurl.com/hollywood2tubes

staff.fwii.uva.nl/p.s.m.mettes

P.S.M.Mettes@uva.nl
Detecting Engagement in Egocentric Video

Yu-Chuan Su and Kristen Grauman
The University of Texas at Austin

1. Engagement in Egocentric Video

Motivation: people do not always engage with what they see and pay different levels of attention to the environment.

Goal: given an egocentric video stream, we want to predict when the camera wearer is engaged with the environment.

Engagement is different from saliency: Previous work [Harel ‘06, Itti ‘09, Rudoy ‘13, …] on visual attention focuses on where the people look but ignores when people are engaged.

2. UT Egocentric Engagement (UT EE) Dataset

We focus our data collection in the following three browsing scenarios:

- Shopping in a Market
- Window Shopping in Mall
- Touring in a Museum

UT EE consists of 27 videos recorded by 9 recorders. The average video length is 30 minutes and the total length is 14 hours.

Engagement Annotation

Frame-level annotation with MTurk. Each video is labeled by 10 Turkers.

3. Data Analysis & Challenges

Annotation consistency

We collect 3 hours of recorder self-annotation to verify the third person annotation.

<table>
<thead>
<tr>
<th></th>
<th>Frame $F_1$</th>
<th>Interval $F_1$</th>
<th>Boundary</th>
<th>Presence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turker vs. Consensus</td>
<td>0.818</td>
<td>0.837</td>
<td>0.914</td>
<td></td>
</tr>
<tr>
<td>vs. Recorder</td>
<td>0.589</td>
<td>0.626</td>
<td>0.813</td>
<td></td>
</tr>
<tr>
<td>Random vs. Consensus</td>
<td>0.426</td>
<td>0.339</td>
<td>0.481</td>
<td></td>
</tr>
<tr>
<td>vs. Recorder</td>
<td>0.399</td>
<td>0.344</td>
<td>0.479</td>
<td></td>
</tr>
</tbody>
</table>

Engagement is predictable from egocentric video!

Challenge of engagement detection
- Diverse visual content
- Being engaged ≠ being static
- Duration of engagement varies significantly

4. Predict Engagement from Motion

A. Estimate frame-wise engagement

B. Generate interval hypotheses

C. Estimate engagement per interval

Temporal pyramid over motion

Interval-level engagement estimator

Final Prediction

5. Experiments

Our method performs the best in all settings and datasets
- Interval hypothesis has clear positive impact
- Appearance feature does not generalize well (UT Ego)
- Saliency/Motion Mag. performs poorly
- We outperform Important region without training on UT Ego

Applications

- VR display
- Camera control

Motivation

People do not always engage with what they see and pay different levels of attention to the environment.

3 hours of recorder self-annotation to verify the third person annotation.

Frame-wise motion descriptor

A. Estimate frame-wise engagement

B. Generate interval hypotheses

C. Estimate engagement per interval

Temporal pyramid over motion

Interval-level engagement estimator

Final Prediction

Applications

- VR display
- Camera control
Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking

**Introduction**

**Discriminative Correlation Filters (DCF):**
- Single-resolution feature map
- Learns a set of discrete filters for target localization
- Outputs discrete detection scores

**Our Approach:**
- Posing the learning problem in the continuous spatial domain
  - Multi-resolution (deep) feature map
  - Learns continuous filters
  - Outputs continuous detection scores

**Advantages**
- Integration of multi-resolution (deep) features
- Accurate sub-pixel (or sub-grid) localization
- Sub-pixel supervision in the learning
- Efficient processing of all available information
- Avoids artefacts caused by explicit resampling

**Applications**
- 1) Object tracking
- 2) Feature point tracking

**Continuous Convolution Operators**

**Interpolation Operator**

\[ J_d : \mathbb{R}^{N_d} \rightarrow L^2(T) \]

\[ J_d \{x^d\}(t) = \sum_{n=0}^{N_d-1} x^d[n] b_d \left( t - \frac{T}{N_d} n \right) \]

**Convolution Operator**

\[ S_f \{x\} = \sum_{d=1}^{D} f^d \ast J_d \{x^d\} \]

\[ g \ast h(t) = \frac{1}{T} \int_0^T g(t-s) h(s) \, ds \]

**Notation**

- \( \hat{g}[k] \) - Fourier coefficients of \( g \in L^2(T) \)
- \( X^d[k] \) - Discrete Fourier transform of \( x^d \)

**Convolution Operator Learning**

**Training loss**

\[ E(f) = \sum_{j=1}^{m} \alpha_j \| S_f \{x^d_j\} - y^j \|^2 + \sum_{d=1}^{D} \| w^d \|^2 \]

\[ \| g \|^2 = \frac{1}{T} \int_0^T |g(t)|^2 \, dt \]

**Fourier Domain**

\[ E(f) = \sum_{j=1}^{m} \alpha_j \left\| \sum_{d=1}^{D} \hat{f}^d \hat{X}^d_j \hat{b}_d - \hat{y}_j \right\|^2 + \sum_{d=1}^{D} \| \hat{w}^d \|^2 \]

Assumption: finitely many non-zero Fourier coefficients.

Gives normal equations: \( (A^H \Gamma A + W^H W) \hat{f} = A^H \Gamma \hat{y} \)

**Object Tracking Framework**

- Features: VGG network (pre-trained on ImageNet)
- Optimization: Conjugate Gradient

**Feature Point Tracking Framework**

Grayscale pixel features

\[ \hat{f}[k] = \sum_{j=1}^{m} \alpha_j X^d_j \hat{b}_d \hat{y}_d \]

Desired continuous output scores (labels)

\[ \sum_{j=1}^{m} \alpha_j X^d_j \hat{b}_d \hat{y}_d = \beta \]

Uniform regularization

\[ w(t) = \beta \]

**Experiments**

**Object Tracking:**

Layer fusion on OTB (100 videos)

<table>
<thead>
<tr>
<th>Layer 0</th>
<th>Layer 1</th>
<th>Layer 2</th>
<th>Layer 3</th>
<th>Layer 4</th>
<th>Layer 5</th>
<th>Layers 0, 1</th>
<th>Layers 0, 2</th>
<th>Layers 1, 5</th>
<th>Layers 0, 1, 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean OP</td>
<td>58.8</td>
<td>78.2</td>
<td>69.0</td>
<td>77.8</td>
<td>70.7</td>
<td>81.8</td>
<td>67.8</td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>AUC</td>
<td>49.9</td>
<td>65.8</td>
<td>51.1</td>
<td>65.7</td>
<td>50.0</td>
<td>67.8</td>
<td>68.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**OTB dataset (100 videos)**

**VOT2016 challenge results** (top 3) [Matej et al., VOT workshop 2016]

<table>
<thead>
<tr>
<th>Tracker</th>
<th>EAO</th>
<th>A</th>
<th>R</th>
<th>A_rank</th>
<th>R_rank</th>
<th>AO</th>
<th>EFO</th>
<th>Impl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. C-COF</td>
<td>0.331</td>
<td>0.530</td>
<td>0.378</td>
<td>12.000</td>
<td>1.000</td>
<td>0.409</td>
<td>0.507</td>
<td>D</td>
</tr>
<tr>
<td>2. TCNN</td>
<td>0.355</td>
<td>0.554</td>
<td>0.268</td>
<td>4.000</td>
<td>3.000</td>
<td>0.485</td>
<td>1.049</td>
<td>S</td>
</tr>
<tr>
<td>3. SSAT</td>
<td>0.321</td>
<td>0.577</td>
<td>0.291</td>
<td>1.000</td>
<td>3.000</td>
<td>0.515</td>
<td>0.475</td>
<td>S</td>
</tr>
</tbody>
</table>

**Feature Point Tracking:**

**The Sintel dataset**
Look-ahead before you leap: end-to-end active recognition by forecasting the effect of motion

Dinesh Jayaraman and Kristen Grauman

UT Austin

Active recognition setting

- passive 1-view setting
- active setting

Components of the active recognition pipeline

- mug?
- bowl?
- pan?

Active recognition setting

- Mug?
- Bowl?
- Pan?

Towards real-world active recognition

- SCENE CATEGORIES
  - SUN 360 panoramas
  - GERMS toy manipulation
  - ModelNet-10 CAD models

- HAND-HELD OBJECTS
  - GERMS toy manipulation [Malmir 2015]

- 3-D CAD MODELS
  - ModelNet-10 CAD models [Wu 2015]

Selected view sequence examples

- SUN 360
- GERMS
- ModelNet-10

Quantitative evaluation results

- SUN 360
- GERMS
- ModelNet-10

Ablation studies

- SUN 360
- GERMS

Our method strongly outperforms representative traditional active recognition approaches on all tasks.

Closely intertwined perception, action and fusion components

Prior art: independent, often heuristic components


Solution: Multi-task training of all active recognition components + auxiliary internally supervised “look-ahead” task.

Joint training

- Perception
- Action selection
- Evidence fusion

End-to-end joint training with gradient descent + REINFORCE

End-to-end joint training with gradient descent + REINFORCE

Our idea

- Not restricted to a single snapshot.
- Strategically acquiring new views.

Complex real-world categories + easily benchmarkable setups.

End-to-end joint training with gradient descent + REINFORCE
Spotlight Sessions
Introduction

Object Recognition Tasks

- (a) image classification
- (b) detection with boxes
- (c) semantic segmentation
- (d) detection with segments

DeepMask for Object Proposals (NIPS15)

- patch
- DeepMask
- mask
- score
- predict mask/score for center object
- sliding window detection

Limitations

- purely feedforward model
- output masks are coarse
- unoptimized architecture

Learning Mask Refinement

Why Top Down Refinement?

- bottom-up network generate coarse semantic mask
- lower layers \(\rightarrow\) spatial information (pixel-level information)
- higher layers \(\rightarrow\) semantic information (object-level information)
- top-down refinement sharpens output using lower-layer features

Refinement Method

- information flow:
  - bottom-up prediction
  - top-down refinement
  - refine offset of pooling
  - double spatial resolution
- general architecture:
  - skip features ++
  - deconv network ++
  - simple, fast, general

\[
M^i+1 = R^i(M^i, F^i)
\]

- \(M^i\) = coarse input mask encoding
- \(F^i\) = corresponding conv features
- \(R^i\) = refinement module \(i\)
- \(M^{i+1}\) = upsampled output mask encoding

Experiments

Object Proposal Results

<table>
<thead>
<tr>
<th></th>
<th>Box Proposals</th>
<th>Segmentation Proposals</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR [50]</td>
<td>AR [75]</td>
</tr>
<tr>
<td>DeepMask</td>
<td>12.6</td>
<td>24.5</td>
</tr>
<tr>
<td>head A</td>
<td>12.0</td>
<td>25.6</td>
</tr>
<tr>
<td>head B</td>
<td>14.0</td>
<td>25.4</td>
</tr>
<tr>
<td>head C</td>
<td>14.4</td>
<td>25.8</td>
</tr>
</tbody>
</table>

DeepMask vs. SharpMask

- novel architecture for object instance segmentation
- state of the art proposal and detection results
- simple, fast, general network for pixel-labeling tasks

Summary

- pedro@opinheiro.com | tl483@cornell.edu
Deep Automatic Portrait Matting

Xiaoyong Shen  Xin Tao  Hongyun Gao  Chao Zhou  Jiaya Jia
The Chinese University of Hong Kong
http://www.cse.cuhk.edu.hk/leojia/projects/automatting

Introduction

Problem

• Image Matting needs tedious user interactions
• The interactions are difficult to meet the algorithm requirement

Challenges

• Learn automatic matting is very difficult
• Rich matte details
• Ambiguous semantic prediction
• Discrepant matte value

Contributions

• We proposed automatic portrait matting
• An end-to-end deep matting CNNs framework
• Novel matting layers
• A matting dataset with 2,000 portraits

Dataset

• 2,000 portraits downloaded from Flickr
• 1,700 for training and 300 for testing
• Different age, gender, pose, hairstyle, background, camera type, etc.
• The matting ground truth is estimated by human well labeled trimap

Our Framework

Trimap Labeling

• Input: RGB image
• Output: trimap representation
• Network: FCN [Long et al. 2015]

Image Matting Layer

• Input: trimap representation
• Output: alpha matte
• Newly-designed layers

Experiments

Results of our deep automatic matting.
Segmentation from Natural Language Expressions
Ronghang Hu1, Marcus Rohrbach1,2, Trevor Darrell1
1University of California, Berkeley 2ICSI, Berkeley

Overview

<table>
<thead>
<tr>
<th>Previous work</th>
<th>This work</th>
</tr>
</thead>
<tbody>
<tr>
<td>category-based semantic segmentation</td>
<td>image segmentation from natural language expressions</td>
</tr>
<tr>
<td>category-based instance segmentation</td>
<td>Grounding visual entities described by referential expressions via dense pixel-wise segmentation:</td>
</tr>
<tr>
<td>bounding box from natural language</td>
<td>- input: image, referential expression</td>
</tr>
<tr>
<td></td>
<td>- output: segmentation mask for the expression</td>
</tr>
<tr>
<td>“a dark horse below a woman in a striped shirt”</td>
<td>“a dark horse below a woman in a striped shirt”</td>
</tr>
</tbody>
</table>

Project page: http://ronghanghu.text objeto

Our Model

Main components:
- Embed the image: spatial feature map through CNN
- Embed the expression: final hidden state in LSTM
- Fully convolutional classification: match input expression to every location on the spatial grid and up sample
End-to-end trainable with back-propagation.

Details model structure

Experiments

Dataset:
- ReferItGame [4] - pixel-wise region annotation for referential expressions

Baseline approaches as comparison:
- Combine per-word segmentation results (bag-of-words)
- Foreground segmentation from bounding box prediction (e.g. [1,3])
- Classification over segmentation proposals (e.g. [5])

Evaluation metric: precision and overall IoU

<table>
<thead>
<tr>
<th>Method</th>
<th>prec@0.5</th>
<th>prec@0.6</th>
<th>prec@0.7</th>
<th>prec@0.8</th>
<th>prec@0.9</th>
<th>overall IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>whole image</td>
<td>5.07%</td>
<td>2.85%</td>
<td>1.58%</td>
<td>0.81%</td>
<td>0.41%</td>
<td>15.12%</td>
</tr>
<tr>
<td>per-word sum</td>
<td>10.97%</td>
<td>5.94%</td>
<td>2.35%</td>
<td>0.45%</td>
<td>0.00%</td>
<td>27.23%</td>
</tr>
<tr>
<td>per-word interaction</td>
<td>9.58%</td>
<td>5.35%</td>
<td>2.20%</td>
<td>0.43%</td>
<td>0.00%</td>
<td>26.69%</td>
</tr>
<tr>
<td>per-word union</td>
<td>10.46%</td>
<td>5.65%</td>
<td>2.28%</td>
<td>0.44%</td>
<td>0.00%</td>
<td>21.72%</td>
</tr>
<tr>
<td>SCOC (1) bbox</td>
<td>11.93%</td>
<td>7.71%</td>
<td>4.33%</td>
<td>1.78%</td>
<td>0.36%</td>
<td>17.84%</td>
</tr>
<tr>
<td>SCOC (1) grabcut</td>
<td>11.93%</td>
<td>7.71%</td>
<td>4.33%</td>
<td>1.78%</td>
<td>0.36%</td>
<td>17.84%</td>
</tr>
<tr>
<td>Groundeld (1) bbox</td>
<td>11.93%</td>
<td>7.71%</td>
<td>4.33%</td>
<td>1.78%</td>
<td>0.36%</td>
<td>17.84%</td>
</tr>
<tr>
<td>Groundeld (1) grabcut</td>
<td>11.93%</td>
<td>7.71%</td>
<td>4.33%</td>
<td>1.78%</td>
<td>0.36%</td>
<td>17.84%</td>
</tr>
<tr>
<td>MCG (5) classification</td>
<td>12.72%</td>
<td>9.88%</td>
<td>7.38%</td>
<td>4.73%</td>
<td>1.88%</td>
<td>18.08%</td>
</tr>
<tr>
<td>Ours (low resolution)</td>
<td>29.54%</td>
<td>21.61%</td>
<td>13.60%</td>
<td>5.94%</td>
<td>0.75%</td>
<td>45.57%</td>
</tr>
<tr>
<td>Ours (high resolution)</td>
<td>34.02%</td>
<td>26.71%</td>
<td>19.32%</td>
<td>11.63%</td>
<td>3.92%</td>
<td>48.03%</td>
</tr>
</tbody>
</table>

References

Semantic Object Parsing with Graph LSTM
Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, Shuicheng Yan
Sun Yat-sen University, 360 AI Institute, National University of Singapore, Adobe Research

Graph LSTM
- Generalize the LSTM for sequential data or multi-dimensional data to general graph-structured data

Graph LSTM Unit
- The hidden and memory states by Graph LSTM can be updated:

\[
g^u_i = \delta(W^u_i x_{i,t+1} + U^u_i h_{i,t} + U^{u*}_i h_{i,t} + b^u),
\]
\[
g^l_i = \delta(W^l_i f_{i,t+1} + U^l_i h_{i,t} + b^l),
\]
\[
g^g_i = \delta(W^g_i f_{i,t+1} + U^g_i h_{i,t} + b^g),
\]
\[
g^c_i = \tanh(W^c_i f_{i,t+1} + U^c_i h_{i,t} + U^{c*}_i h_{i,t} + b^c),
\]
\[
m_{i,t+1} = \sum_{j \in N_G(i)} (\mathbf{1}(q_j = 1)g^l_{ij} \odot m_{j,t+1} + \mathbf{1}(q_j = 0)g^g_{ij} \odot m_{j,t}) + g^c_{ij} \odot m_{i,t} + g^u_{ij} \odot g^c_{ij},
\]
\[
h_{i,t+1} = \tanh(g^c_{ij} \odot m_{i,t+1}).
\]

Network Architecture for Semantic Object Parsing
- The Graph LSTM layers are stacked to sequentially update the hidden states of all super-pixel nodes.

- Graph LSTM vs locally fixed factorized LSTM
  Using richer and adaptive local contexts (i.e., number of neighbors) to update the states of each pixel can lead to better parsing performance.

- Adaptive forget gates vs Identical forget gates
  Diverse semantic correlations with local context can be considered and treated differently during the node updating.

- Confidence-driven node updating scheme
  The features of superpixel nodes with higher foreground confidences embed more accurate semantic meanings and thus lead to more reliable global reasoning.
SSHMT: Semi-supervised Hierarchical Merge Tree for Electron Microscopy Image Segmentation

Ting Liu1, Xiaomiao Zhang2, Mehran Javanmardi1, Nisha Ramesh3, Tolga Tasdizen1
1 Scientific Computing and Imaging Institute, University of Utah
{ting, mehran, nshramesh, toliga}@sci.utah.edu
2 CSAIL, MIT
miao86@mit.edu

Summary
• Quick background: region-based methods after pixel-based membrane detection accuracy for neuron segmentation in electron microscopy (EM) images
• Goal: learn scoring function for region merging with few training samples
• Contributions: extend existing hierarchical merge tree (HMT) framework with:
  • An unsupervised loss term that enforces structural consistent predictions about region merging
  • A Bayesian model for probabilistic semi-supervised learning and automatic parameter estimation
• Result: with 3% to 7% of labeled data, consistent performance close to supervised HMT with full labeled sets

Hierarchical Merge Tree
• Start with over-segmenting initial superpixels, build a merge tree
• Learn to score potential merges: boundary classifier
• Infer final segmentation

Unsupervised Constraint
• Consistency constraint: no merge after split in leaf-to-root path
• In disjunctive normal form: let \( f_i \) be merge indicator
  • \( F = \bigvee_{i=1}^{n} \bigwedge_{j=1}^{m} \bigvee_{k=1}^{p} \neg f_j \) is true iff consistent
  • \( \mathbf{F} = (-f_0 \wedge f_0) \lor (f_0 \wedge \neg f_0) \lor (\neg f_0 \wedge f_0) \) for Fig. 1(b)
• Real-value approximation: let \( f_i \in [0, 1] \) be boundary classifier prediction
  • \( \bar{F} = 1 - \prod_{j=1}^{m} \left( 1 - f_j \cdot \Pi_{k=1}^{p} (1 - f_k) \right) \) = 1 iff consistent

Semi-supervised HMT
• Goal: learn a boundary classifier whose predictions minimize \( \|1 - \bar{F}\| \) for any tree path
• Boundary classifier: \( f_w(x) = 1/\left(1 + \exp(-w^T x)\right) \)
• Bayesian model to learn \( f_w \):

\[
P(w|X, y, \alpha_u, \sigma_u, \alpha_s, \sigma_s) \propto P(w) \cdot P(1|X, w, \alpha_u) \cdot P(y|X, w, \sigma_s) \propto \exp\left(\frac{-\|w\|^2}{2}\right)^{\frac{1}{2}} \frac{1}{(\sqrt{2\pi}\sigma_u)} \exp\left(-\frac{\|1 - \bar{F}_w\|^2}{2\sigma_u^2}\right) \frac{1}{(\sqrt{2\pi}\sigma_s)} \exp\left(-\frac{\|y - \bar{F}_w\|^2}{2\sigma_s^2}\right)
\]

• Objective function:

\[
J(w, \alpha_u, \sigma_u) = \frac{1}{2} \|w\|^2 + \frac{1}{2\sigma_u^2} \|1 - \bar{F}_w\|^2 + N_u \log \sigma_u + \frac{1}{2\sigma_s^2} \|y - \bar{F}_w\|^2 + N_s \log \sigma_s
\]

• Gradient descent on \( w \); alternately estimate \( \sigma_u \) and \( \sigma_s \) using closed-form solutions to \( \partial J/\partial \sigma_u = 0 \) and \( \partial J/\partial \sigma_s = 0 \)
• Infer final segmentation as in HMT

Results
• Datasets: mouse neuropil (2D SBFSEM, 7000×7000×70), mouse cortex (3D S^5EM, 1024×1024×100), and Drosophila (3D FIBSEM, 500×500×500)
• Randomly subsample labeled data for as supervised data; all samples as unsupervised data; repeat 50 times whenever possible
• Evaluation metric: adapted Rand error (lower is better)
• Quantitative results are shown in Fig. 2; qualitative results are shown in Fig. 3 and 4.
The Curious Robot: Learning Visual Representations via Physical Interactions

Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park and Abhinav Gupta

**Summary**

Physical Interaction Data

Learned Visual Representation

**Our Approach**

Shared Representation Learning

**Visual Representations**

Nearest Neighbors

Neuron Activations

**Classification**

Accuracy (%)

- Random Initialization: Household 25.0, UW RGBD 46.8, Cailtech-256 24.2
- Pretrained on robot tasks: Household 35.4, UW RGBD 69.3, Cailtech-256 31.7
- AlexNet pretrained: Household 62.5, UW RGBD 82.0, Cailtech-256 65.6
- Autoencoder: Household 29.6, UW RGBD 65.7, Cailtech-256 28.0
- Only Invariance: Household 31.5, UW RGBD 66.0, Cailtech-256 25.2

**Image Retrieval**

- Category level recall@k
  - RandomNet: 15.0, 46.6, 65.2, 80.0
  - alexNet: 85.4, 95.3, 96.9, 98.9
  - ourNet: 83.3, 91.8, 94.6, 96.6

- Instance level recall@k
  - RandomNet: 6.2, 21.9, 33.1, 47.5
  - alexNet: 68.6, 85.7, 90.3, 94.1
  - ourNet: 72.0, 83.1, 87.5, 90.9

**Related Work**

Examples:

- Single Images
  - Doersch et al. 2015
  - Agrawal et al. 2015

- Videos
  - Wang et al. 2015

- Sound
  - Owens et al. 2016

**Ablation Analysis**

- All robot tasks: Household 35.4, UW RGBD 69.3, Cailtech-256 31.7
- Except Grasp: Household 30.9, UW RGBD 63.2, Cailtech-256 26.3
- Except Push: Household 35.6, UW RGBD 71.0, Cailtech-256 27.9
- Except Poke: Household 34.2, UW RGBD 68.4, Cailtech-256 28.9
- Except Invariance: Household 32.4, UW RGBD 71.1, Cailtech-256 29.7
Image Co-localization by Mimicking a Good Detector’s Confidence Score Distribution

Yao Li Lingqiao Liu Chunhua Shen Anton van den Hengel
The University of Adelaide, Australia

Problem definition
The image co-localization problem:
• Input: a set of images with one common object.
• Output: localized bounding boxes of each instance in each image.

Learning detectors by modeling detection score distribution

Some notations:
• Input images: \( I = \{I_1, I_2, \ldots, I_n\} \).
• Object proposals for each image \( I_i: B_i = \{B_{i1}, B_{i2}, \ldots, B_{iM}\} \).
• Feature of each object proposal \( B_{ij} \in B_i; \phi(B_{ij}) \in \mathbb{R}^k \).

Assuming there is a learnt detector parametrized by \( w \in \mathbb{R}^k \) and \( b \in \mathbb{R}^1 \), the particular detection confidence score of a proposal is defined as follows:

\[
s_{ij} = f(w^T \phi(B_{ij}) + b),
\]

where \( f() \) is the softplus function. The scores \( s_{ij} \) in each image are further normalized into \( p_i = [p_{i1}, p_{i2}, \ldots, p_{iM}]^T \).

The objective for learning the common object detector:

\[
\min_{w, b} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} L(p_i) + \lambda \|w\|_2^2,
\]

where the loss function \( L(p_i) \) is the key ingredient of this work.

As shown in the motivation, we hope the \( p_i \) to be a sparse vector. In this work, we utilize the Shannon entropy as a sparsity indicator,

\[
L(p_i) = -\sum_{j=1}^{M} p_{ij} \log p_{ij}.
\]

So the optimal values of the weight and bias of the detector are given by:

\[
w^*, b^* = \arg \min_{w, b} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} \log p_{ij} + \lambda \|w\|_2^2.
\]

We use stochastic gradient descent (SGD) for minimizing the objective function.

Thus, for the purpose of localizing common objects, we incorporate the objectness score \( o_i \) of each proposal \( B_{ij} \) into Eq.(1)

\[
s_{ij} = o_i f(w^T \phi(B_{ij}) + b).
\]

Minimizing Eq. (4) using \( s_{ij} \) defined in Eq. (5) gives a stable solution regardless of initialization.

Motivation
Our motivation is from the detection score distribution of a strongly-supervised object detector when applied to an image.

Key observation. Only a small minority of proposals are given high detection confidence scores while most of them are associated with low scores.

Experiments
• Implementation details: CNN feature, Edgebox.
• Datasets: PASCAL VOC 07 & 12, six ImageNet categories.
• Evaluation metric: CorLoc (IoU over 50%).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC 2007</td>
<td>24.0</td>
<td>36.6</td>
<td>44.0</td>
</tr>
<tr>
<td>Ours</td>
<td>43.8</td>
<td>43.8</td>
<td>48.8</td>
</tr>
</tbody>
</table>

Table : VOC 2007

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>VOC 2012</td>
<td>24.0</td>
<td>42.4</td>
<td>37.7</td>
</tr>
<tr>
<td>Ours</td>
<td>48.5</td>
<td>48.5</td>
<td>48.5</td>
</tr>
</tbody>
</table>

Table : VOC 2012

Table : six ImageNet categories

Figure : From top to bottom: input images (predicted boxes in red), detection heat maps, segmentation results.
Facilitating and Exploring Planar Homogeneous Texture for Indoor Scene Understanding
Shahzor Ahmad and Loong-Fah Cheong
National University of Singapore

Motivation
- Ubiquity of homogeneous texture in man-made environments
- Inadequacy of existing tools to tap its potential for high-level scene understanding
  - outliers
  - perspective distortion
  - photometric severities

Texture Frequency Projection Model
- Estimate variable frequency in image and minimize reprojection error over constant frequency on intermediate plane and $p_i$.

Dominant Frequency Estimation via Graph Cuts Optimization (GCO)
- Unary term favours Gabor filter $f_x = (\Omega_x, \theta_x) \in \mathcal{C}$ that maximizes response (same as DEMOD [1]):
  \[ P_r(f_x) = \frac{\alpha}{A(f_x)p} \]
- Pairwise term smooths radial frequency and orientation:
  \[ V(f_x, f_y) = \beta (\Omega_x - \Omega_y)^2 + (\sin\theta_x - \sin\theta_y)^2 + (\cos\theta_x - \cos\theta_y)^2 \]
- GCO resolves frequency drift due to outliers:
  - $GCO$ resolves quadrant ambiguity:

Affine Rectification
- $N = 30$ cropped textures from MIT Indoor67 [2]
- Mean Estimation Error:
  \[ \sum_{n=0}^{N-1} \sqrt{(\hat{h}_n - h_n)^2 + (\hat{b}_n - b_n)^2} \]

Scene Geometric Layout Estimation
- Existing work [5] is fraught with challenges:
  - > 3 dominant planar directions (b, f, g)
  - absence of straight lines in a direction (c)
  - non-Manhattan structure (e)
  - Incorrect room face localization (a, h)
- Proposed approach does not rely upon vanishing points or machine learning to localize faces

Detection & Geometric Layout: Quantitative Evaluation
- 300 MIT Indoor67 [2] images manually annotated (left)
- Precision and recall computed based on estimated geometric class for detections (right)
- AP: 0.53 (ours) vs. 0.15 (TILT [3])

Scene Classification
- Rectification aligns features to a canonical coordinate frame, mitigating in-class variance and improving recognition
- Features extracted upon rectification are class-discriminative and complementary to original descriptors
- All state-of-the-art methods on MIT Indoor 67 [2] employ machine learning for feature extraction
- Proposed approach is entirely handcrafted, achieving 64.54% when combining SIFT, CENTRIST & HOG descriptors

References
An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild

Wei-Lun Chao*1, Soravit Changpinyo*1, Boqing Gong2, and Fei Sha1,3
1U. of Southern California, 2U. of Central Florida, 3U. of California, Los Angeles

Highlights
- Study generalized zero-shot learning (GZSL)
- Test data & possible labels from BOTH Seen + Unseen classes, not just from Unseen ones.
- Propose an effective calibration method to adapt ZSL algorithms to perform well in GZSL
- Develop a metric AUSUC for GZSL evaluation
- Establish a performance upper bound of GZSL via idealized semantic embeddings

ZSL algorithms in GZSL setting
- Joint labeling space of Seen (S) and Unseen (U):
  \( \mathcal{T} = S \cup U \)
- Scoring function for each class
  - DAP [Lampert et al., CVPR 09]:
    \( f_u(x) = w(a_u)^T x \)
  - ConSE [Norouzi et al., ICLR 14]:
    \( f_u(x) = \cos(s(x), a_u) \)
  - SynC [Changpinyo et al., CVPR 16]:
    \( f_u(x) = P(a_u|x) \)
- Classification by Direct Stacking
  \( \hat{y} = \arg \max_{c \in \mathcal{T}} f_c(x) \)

Experiments & Analysis
- Datasets (|S|/|U|): AwA (40/10), CUB (150/50), ImageNet (1,000/20,842)
- Semantic embeddings: attributes for AwA/CUB, word vectors for ImageNet
- Visual features: 1,024-dim GoogLeNet features
- Evaluation: AUSUC on (class-normalized) classification accuracy or Flat Hit@K

ZSL vs. Generalized ZSL
- Seen classes come with labeled examples. Unseen classes come without.
- Goal: Expand classifiers and label space from Seen classes to Unseen ones = dealing with long-tailed object distributions and recognition in the wild.
- Relate Seen and Unseen classes with Semantic embeddings (attributes, word vectors, etc.)
- Training: Learn from Seen classes’ images and semantic embeddings
- Testing: (Conventional) Zero-Shot Learning (ZSL)
  Classifying images from Unseen into the label space of Unseen
  Generalized Zero-Shot Learning (GZSL)
  Classify images from BOTH Seen + Unseen into the label space of BOTH Seen + Unseen
  Much more challenging!

Proposed Calibration Method & Metric
- Classification by Calibrated Stacking
  \( \hat{y} = \arg \max_{c \in \mathcal{T}} f_c(x) - \gamma(I[c \in S]) \)
- Area Under Seen Unseen accuracy Curve (AUSUC)
  Varying the calibration factor leads to Seen-Unseen Accuracy Curve (SUC) of \( (A_{\gamma \rightarrow \tau}, A_{S \rightarrow \tau}) \)
- Area Under SUC (AUSUC) as the metric for GZSL

Which ZSL method is more robust to GZSL?

How far are we from the ideal multi-class & GZSL performance?

Analysis on ImageNet-2K: K = 1
- Multi-class classifiers trained on data from S & U
- Idealized semantic embeddings (G-attr) = Average of visual features for each class

Analysis on ImageNet All:
- Flat hit@K
- WORD2VEC: 0.006/0.034
- G-attr from 1 image: 0.018/0.071
- G-attr from all images: 0.067/0.236

From Derek Hoiem’s slides
## Dataset Overview

**ReCoCo**:
1. giraffe on left
2. 1st giraffe on left
3. giraffe head down
4. giraffe hugging another giraffe

**ReCoCo**:  
- Test A: 49.84%
- Test B: 4.53%
- Validation: 43.65%
- 11.73%

**RefCOCO**:  
- Test A: 70.25%
- Test B: 0.578

**RefCOCO+**:  
- Test A: 0.331
- Test B: 0.458

**RefCOCOg**:  
- Test A: 28.67%
- Test B: 0.140

### Two Tasks

**Task 1: Comprehension**
Which object is "Girl on the left" indicating?

**Task 2: Expression Generation**
Generate referring expression for this target person.

## Results

### Comprehension results on the ReCoCo, ReCoCo+, and ReCoCoG datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Set</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Rouge-L</th>
<th>Meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReCoCo</td>
<td>Test A</td>
<td>0.196</td>
<td>0.322</td>
<td>0.440</td>
<td>0.179</td>
<td>0.214</td>
</tr>
<tr>
<td>ReCoCo+</td>
<td>Test A</td>
<td>0.183</td>
<td>0.310</td>
<td>0.430</td>
<td>0.175</td>
<td>0.208</td>
</tr>
<tr>
<td>ReCoCoG</td>
<td>Test A</td>
<td>0.165</td>
<td>0.298</td>
<td>0.401</td>
<td>0.162</td>
<td>0.194</td>
</tr>
</tbody>
</table>

### Expression Generation Results: Bleu, Rouge, Meteor evaluations for ReCoCo and ReCoCo+.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Set</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Rouge-L</th>
<th>Meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReCoCo</td>
<td>Test A</td>
<td>0.340</td>
<td>0.518</td>
<td>0.646</td>
<td>0.324</td>
<td>0.441</td>
</tr>
<tr>
<td>ReCoCo+</td>
<td>Test A</td>
<td>0.325</td>
<td>0.507</td>
<td>0.636</td>
<td>0.320</td>
<td>0.437</td>
</tr>
</tbody>
</table>

### Human Evaluations on expression generation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Set</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Rouge-L</th>
<th>Meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReCoCo</td>
<td>Test A</td>
<td>0.332</td>
<td>0.507</td>
<td>0.636</td>
<td>0.320</td>
<td>0.437</td>
</tr>
<tr>
<td>ReCoCo+</td>
<td>Test A</td>
<td>0.325</td>
<td>0.507</td>
<td>0.636</td>
<td>0.320</td>
<td>0.437</td>
</tr>
</tbody>
</table>

### Fraction of images with duplicate expressions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test Set</th>
<th>Bleu 1</th>
<th>Bleu 2</th>
<th>Bleu 3</th>
<th>Rouge-L</th>
<th>Meteor</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReCoCo</td>
<td>Test A</td>
<td>0.325</td>
<td>0.507</td>
<td>0.636</td>
<td>0.320</td>
<td>0.437</td>
</tr>
<tr>
<td>ReCoCo+</td>
<td>Test A</td>
<td>0.325</td>
<td>0.507</td>
<td>0.636</td>
<td>0.320</td>
<td>0.437</td>
</tr>
</tbody>
</table>

### Loss function

\[
L = - \sum \log P(r_t|v_t) - \lambda \max(0, M - \log P(r_t|v_t)) - \log P(r_t|v_t))
\]

*Generation loss, Max-margin loss or Maximum Mutual Information (MM)*
Binary Hashing for Image Search

- Binary hashing
  - Efficient storage and fast searching
  - Attractive approach for large scale visual search
- Traditional approach for hashing
  - Minimizing the specific loss function under the binary constraint on the codes
  - Coupling of the hash function and the binary constraint → very challenging to solve
- Two-step approach for hashing [TSH[15], FastHash[16]]
  - Inference binary codes
  - Learn hash function, given binary codes
  - Reduce the complexity of the coupled problem, flexible using of different hash functions

Contributions

- Unified formulation for both supervised and unsupervised hashing
- Two novel approaches for inferring binary codes
  - Semidefinite Relaxation (SDR)
  - Augmented Lagrangian (AL)

Unified formulation

- Input: $S$: similarity matrix between samples, i.e., pairwise distance matrix for unsupervised or pairwise label matrix for supervised; $L$ : code length; $n$: number of training samples
- Target: learn binary codes $Z$ s.t. the similarity matrix $S$ is preserved in Hamming space, i.e., solving

$$\min_{Z \in \{-1,1\}^n} \frac{1}{2} \|Z^T S - Z\|_F^2$$

- Supervised: $Y = L - 2S$, where $S$ is pairwise distance matrix
- Unsupervised: $Y = L - 2S$, where $S$ is label pairwise matrix

Using coordinate descent approach for solving above NP-hard, i.e., solving one row of $Z$ at a time. Let $x = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$, we solve RQP

$$\min_k x^T A x$$
$$s.t. \text{diag}(X) = 1, X \succeq 0, \text{rank}(X) = 1$$

Semidefinite Relaxation (SDR) approach

Let $B = A - \lambda I$, where $\lambda_i$ is the largest eigenvalue of $A$; $X = xx^T$ 

→ solve equivalent problem

$$\min_k \text{trace}(BX)$$
$$s.t. \text{diag}(X) = 1, X \succeq 0$$

Solving: Using Convex OPT packages: SeDuMi, SDPT3 → achieving the global optimal solution $X^*$

- Recover binary solution $k$ from $X^*$
- Solving: apply randomized rounding process several times and select the best solution
  - Generate $l$ by $\xi \sim N(0, X^*)$
  - Get feasible point: $x = \text{sgn}(\xi)$

Bound on objective value at $k^*$: Let $f_{opt}$ be global optimum objective value of (3) and $f_{SDR-conv}$ = $x^T B x$, we have

$$f_{opt} \leq E[f_{SDR-conv}] \leq \frac{2}{\lambda_{max}} f_{opt}$$

Augmented Lagrangian (AL) approach

Let $\theta(k) = [(x_1^2 - 1), \ldots, (x_n^2 - 1)]^T \lambda = [\lambda_1, \ldots, \lambda_n]^T$: Lagrange multipliers, by using AL for (2), we minimize the unconstrained augmented Lagrangian function

$$L(x; \lambda) = x^T A x - \lambda^T \theta(k) + \frac{1}{2} \|\theta(k)\|^2$$

- When $\mu$ is large → penalize the binary constraint violation severity → force the minimizer of the AL function (7) closer to the feasible region of the original problem (2)
- Theoretically, not necessary to take $\mu \rightarrow \infty$ in order to achieve a local optimum of (2)

Compare SDR and AL approaches

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CIFAR10</th>
<th>MNIST</th>
<th>SUN397</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(c)</td>
</tr>
<tr>
<td>CIFAR10</td>
<td><img src="cifar10.png" alt="Image" /></td>
<td><img src="mnist.png" alt="Image" /></td>
<td><img src="sunset.png" alt="Image" /></td>
</tr>
<tr>
<td>MNIST</td>
<td><img src="cifar10.png" alt="Image" /></td>
<td><img src="mnist.png" alt="Image" /></td>
<td><img src="sunset.png" alt="Image" /></td>
</tr>
<tr>
<td>SUN397</td>
<td><img src="cifar10.png" alt="Image" /></td>
<td><img src="mnist.png" alt="Image" /></td>
<td><img src="sunset.png" alt="Image" /></td>
</tr>
</tbody>
</table>

Evaluation of Supervised Hashing

- Two novel approaches, i.e. Semidefinite Relaxation and Augmented Lagrangian, for binary code inference
- SDR slightly outperforms AL
- SDR and AL outperform most of compared methods.

Evaluation of Unsupervised Hashing

- SDR slightly outperforms AL
- SDR and AL outperform most of compared methods

Conclusion

- Unified framework for both unsupervised and supervised hashing
- Two novel approaches, i.e. Semidefinite Relaxation and Augmented Lagrangian, for binary code inference
Introduction

Dense-CRF:
- Popular framework for vision problem modelisation.
- MAP estimation problem is of main interest.
- Allow modelisation of long range interactions.

Main challenge:
- Methods for sparse CRF does not scale for dense ones.
- Implementation requires $O(n^2)$ computations.

Related work

MF based methods
Krahenbuhl and Koltun initial work [Kra. 2011]
- Use MF to approximate the MAP problem.
- Restricted to Gaussian pairwise term.
- Use Permutohedral Lattice for $O(1)$ computations.

SDP based method
[Boye 2015]
- SDP relaxation of the MAP problem.
- Use low-rank approximation to make it tractable.

Preliminaries

Energy function
$$E(x) = \sum_{a} \psi_a(x_a) + \sum_{a \neq b} \psi_{ab}(x_a, x_b).$$

Permutohedral Lattice
[Alban Desmaison 2015]
- Allows to compute the following in $O(n)$ complexity (instead of $O(n^2)$):
  $$\psi_a(x_a) = \sum_{i=1}^{n} \psi(x_a[i], a).$$

Motivation

For sparse CRFs, relaxation-based methods are both efficient and have strong guarantees. But these methods does not scale with the number of pairwise connections. Thus the only feasible algorithm for dense CRF MAP estimation is the MF-based approximation. Here we present efficient algorithms to solve convex relaxations of the original MAP estimation problem. We thus present the first efficient algorithms with guarantees for the MAP estimation problem for Dense CRFs.

QP relaxations

Convex QP relaxation
- Conditional gradient and optimal step size in $O(n)$.
- Convex problem not sensible to initialisation.

Generic difference of convex relaxation
- Use concave convex procedure iteratively.
- Each step of the CCCP requires solving a convex QP as above.
- Monotonic decrease to a local minima.

Specialised difference of convex
- Works only for negative semi-definite pairwise potentials.
- Use concave convex procedure iteratively.
- The convex QP is very easy to solve.
- Fast monotonic decrease to a local minima.

LP relaxation

LP relaxation of the MAP estimation problem
[Kleinberg 2002]
- Best possible relaxation of the original problem.
- Tight in the 2 label case.
- 2-approximate for more than 2 labels.
- Can compute the subgradient in $O(n \log(n))$.
- Use a divide and conquer approach based on the permutohedral lattice algorithm.
- Do not compute full matrix vector product.
- Consider triangular matrix vector product.
- Performs divide and conquer to always consider full blocks.
- Reach the unique optimal solution via projected subgradient descent.

Results

The convex relaxation-based methods presented here reliably lead to better energy compared to the MF-based approach.

Stereo Matching
Evolution of the energy with respect to time for the Teddy image and the final matchings:

Semantic Segmentation

<table>
<thead>
<tr>
<th>Original Image</th>
<th>Ground Truth</th>
<th>MF Parameters</th>
<th>DCneg Parameters</th>
<th>QPneg Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image16.png" alt="Image" /></td>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Complexity of Discrete Energy Minimization Problems
Mengtian Li\textsuperscript{1} Alexander Shekhovtsov\textsuperscript{2} Daniel Huber\textsuperscript{1}
Carnegie Mellon University\textsuperscript{1} Graz University of Technology\textsuperscript{2}

Abstract

Energy minimization problems vary greatly in approximability

• Energy minimization is NP-hard \& is it approximable? Not yet resolved
• Sometimes yes: Potts, Metric, Logic MRF

We prove that QPBO, planar energy with 3+ labels, and general energy minimization are all inapproximable

• Useful for algorithm design — finding “good” subclasses
• In practice, useful for model selection

Complexity Axis & Main Results

Theorem:

- QPBO (binary labels) is complete in exp-APX.
- General energy minimization is complete in exp-APX.
- Planar energy with 3+ labels is complete in exp-APX.

Optimization & Approximation

- Optimization problems
  - Approximation ratio $f(x)/f(x^*)$, $f(x^*) > 0$
  - APX — constant ratio approximation
  - F-APX — approximation ratio is a function of class $F$ of the input bit length
- Relations of complexity classes
  \[ \text{PO} \subseteq \text{APX} \subseteq \log \text{-APX} \subseteq \text{poly-APX} \subseteq \text{exp-APX} \subseteq \text{NPO} \]

Details

Non-deterministic Polynomial time Optimization (NPO)
- The set of instances is recognizable in polynomial time
- The solution’s feasibility is verifiable in polynomial time
- A positive objective value

Polynomial time Optimization (PO)
- The problem is in NPO, and it is solvable in polynomial time

Approximation-Preserving reduction (AP-reduction)
- Reduce NPO problem $P_1$ to another NPO problem $P_2$

\[ x_1 : \text{instance of } P_1 \rightarrow \pi(x_1) : \text{instance of } P_2 \]

\[ y_1 : \text{solution of } P_1 \rightarrow \sigma(y_1) : \text{solution of } P_2 \]

For a given positive constant $\alpha$, the mappings must satisfy

\[ \frac{f(\pi(x_1))}{f(y_1)} \leq \alpha \quad \text{implies} \quad \frac{f(\sigma(y_1))}{f(y_1)} \leq \alpha + 1 \]

- \[ \mathcal{C}-\text{hard} \& \mathcal{C}-\text{complete} \]
  - A problem is \[ \mathcal{C}-\text{hard} \] if any problem in complexity class $\mathcal{C}$ can be reduced to it
  - A $\mathcal{C}$-hard problem is \[ \mathcal{C}-\text{complete} \] if it belongs to $\mathcal{C}$
  - Intuitively, a complexity class $\mathcal{C}$ specifies the upper bound on the hardness of the problems within, $\mathcal{C}$-hard specifies the lower bound, and $\mathcal{C}$-complete exactly specifies the hardness

Problem W3SAT-triv

INSTANCE: Boolean CNF formula $F$ with variables $x_0, …, x_n$, and each clause assuming exactly 3 variables; non-negative integer weights $w_{x_0}, …, w_{x_n}$

SOLUTION: Truth assignment $r$ to the variables that either satisfies $F$ or assigns the trivial, all-true assignment

\[ \text{MEASURE: } \min \sum_{i=0}^{n} w_{x_i} r(x_i) \]

References

A Convex Solution to Spatially-Regularized Correspondence Problems

Thomas Windheuser and Daniel Cremers
Technische Universität München

Spatially-Regularized Correspondence Problem

We show there is a deep geometric structure underlying the spatially-regularized correspondence problem that is essential for solving the problem in a convex fashion. To the best of our knowledge we are the first that introduce this structure for image correspondence problems.

Main Contribution

The set of $\mathcal{J}$-vectors $\mathcal{J} \Gamma$ is a vector space.

Simple $\mathcal{J}$-Vectors

- simple $\mathcal{J}$-vectors have the form $\mathcal{J}_1 \mathcal{J}_2 \in \mathcal{J} \Gamma^{2,1}$
- equivalence between simple $\mathcal{J}$-vectors and oriented 2-dimensional subspaces plus some positive area
- $\mathcal{J} \Gamma^{2,1} = \{ \mathcal{J}_1 \mathcal{J}_2 \}$

Non-Simple $\mathcal{J}$-Vectors

- $\mathcal{J}_1 \mathcal{J}_2 + \mathcal{J}_3 \mathcal{J}_4$ is non-simple
- $[\mathcal{J}_1 \mathcal{J}_2 \mathcal{J}_3 \mathcal{J}_4] \neq 0$, but $[\mathcal{J}_1 \mathcal{J}_2 + \mathcal{J}_3 \mathcal{J}_4] = \sqrt{2}$
- Solution: the mass norm

Convex Optimization Problem

- make coefficients of $\omega$ differentiable
- replace $\mathcal{H}$ by $\mathcal{H}^1$
- $\mathcal{H}^{1} \wedge \omega = 0 \Rightarrow \text{div} \omega = 0$, where $\text{div} \omega = \sum_{i=1}^n \frac{\partial \theta_i}{\partial x_i}$

Discrete Optimization Problem

- $\min c(\mathcal{P}) \sum |X_j|$
- s.t. div $\omega = 0$ and $\pi_1(\omega) = \pi_2(\omega) = 1$
Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks
Chuan Li and Michael Wand

**Objectives**
- Speed up DNNs based texture synthesis
- Preserve high quality

**Challenges**
- Learn textures with Generative Adversarial Networks
- Guided and Un-guided synthesis
- Generalize to new inputs (image/video/noise)

**Markovian GANs**

**Deconvolution**

**Feed-forward**

**Results**

**Visualization**

**Comparison**

Flickr user Steve K
From the coarse level $l = L - 1$, we upsample the signal by a factor of 2 at each level by solving the following weighted least square (WLS) using a recent FGS solver.

Guided interp.: Next, another WLS is solved with the output $d^*$ as guidance and bicubic interpolated signal as input.

Joint filtering: We then check the consistency between the output and the bicubic upsampled data and pick the most consistent points to add to the data mask map. The process is repeated until the finest level ($l = 0$).

The common objective is to densify a set of sparse data, either regularly distributed or scattered, to a full image grid via 2D guided interpolation.

**Overview**

Depth upsampling and motion interpolation are required to generate a dense, high-quality, and high resolution depth map or optical flow field.

**Challenges:** (also loss of details/boundaries, complex, slow...)
- Texture-copying artifacts due to inconsistent structures
- Large occlusions, long-range propagation, extrapolation

Our solution: A hierarchical (coarse-to-fine), multi-pass guided interpolation framework that divides the problems into a sequence of interpolation tasks each with smaller scale factors

**Pipeline**

From the coarse level $l = L - 1$, we upsample the signal by a factor of 2 at each level by solving the following weighted least square (WLS) using a recent FGS solver. Guided interp.: $\varepsilon(d_i) = (d_i - d_n)^T M(d_i - d_n) + \lambda d_i^T A d_i$.

Next, another WLS is solved with the output $d_i$ as guidance and bicubic interpolated signal as input. Joint filtering: $\varepsilon(d_i) = (d_i - d_n)^T M(d_i - d_n) + \lambda d_i^T A d_i$.

We then check the consistency between the output and the bicubic upsampling data and pick the most consistent points to add to the data mask map $\tilde{m}_i$. The process is repeated until the finest level ($l = 0$).

**Results**

1000x faster than AR

Average Depth Upsampling Error on ToF Synthetic Dataset (6 cases)

- **Depth Upsampling**
  - **Runtime (sec) to upsample a 272 x 344 depth map to 1088 x 1376**
    - MRF+nlm: 170, TGV: 420, AR: 900, GF: 1.3, CLMF: 2.4, FGI (ours): 0.6
    - **1000x faster than AR**

- **Motion Interpolation**
  - Performance (EPE) on MPI Sintel training set
    - WLS, EpicFlow-NW, EpicFlow-LA, FGI (ours)
    - Clean: 3.23, 3.17, 2.65, 2.75, Close to AR
    - Final: 4.68, 4.55, 4.10, 4.14, over 2x faster
    - Runtime (sec): 0.21, 0.80, 0.94, 0.39

- **Performance (EPE) on MPI Sintel testing benchmark**
  - Close to AR
  - 650x faster than TGV

**Remark**
- A generic unified approach for guided depth and motion interpolation
- Generally applicable to other edge-aware filters e.g. the guided filter
- Competitive results while running much faster than task-specific methods
- Simple & effective: no variational minimization/extra DT/edge detection
Learning High-Order Filters for Efficient Blind Deconvolution of Document Photographs

Lei Xiao
University of British Columbia

Jue Wang
Adobe Research

Wolfgang Heidrich
KAUST

Michael Hirsch
MPI for Intelligent Systems

Introduction
• Taking photographs of text documents, instead of scanning them, has become increasingly common due to the popularity of mobile cameras.

• However, photos taken by hand-held cameras are likely to suffer from blur caused by camera shake during exposure.

• This is critical for document images, as slight blur can make Optical Character Recognition (OCR) fail.

• Observation: common font-size document images are dominated by high-order structures, making traditional sparse gradient priors ineffective.

Contribution
• Demonstrate the importance of using high-order filters in text document image restoration.

• Propose a new algorithm for fast and high-quality deblurring of document photographs, suitable for processing high resolution images captured by modern cameras.

• Robust to page orientation, font style and text language, even though such variants are not included in our training dataset.

Algorithm
• Basic idea: use high-order filter priors learned via discriminative learning

\[ x_k = \arg\min_{x_k} \| b - k \otimes x \|^2 + \sum_{i=1}^N \rho_i(F_i(x)) + \tau \| k \|_1, \quad \text{s.t.} \quad k \geq 0, \| k \|_1 = 1 \]

• Framework: multi-scale, interleaved shrinkage-fields network

\[
\begin{align*}
\text{latent image } x_t & : \quad x_t = F^{-1} \left( \frac{1}{z_t} \sum_{i=1}^N \left( F_i(x_{t-1}) \right)^2 + \mu \right) \\
\text{(Shrinkage-Field model, CVPR14) \quad \text{Training:}} \quad & \left( \lambda_1, \lambda_2, \alpha \right) = \arg\min_{x, \lambda, \alpha} \| x - x_{t-1} \|_2^2 \\
\text{secondary image } z_t & : \quad z_t = F^{-1} \left( \frac{1}{k_t} \sum_{i=1}^N \left( F_i(z_{t-1}) \right)^2 + \mu \right) \\
\text{(Shrinkage-Field)} \quad \text{Training:} \quad & \left( \lambda_1, \lambda_2, \alpha \right) = \arg\min_{x, \lambda, \alpha} \| x - z_{t-1} \|_2^2 \end{align*}
\]

• Limitant image update:

\[ x_t = F^{-1} \left( \frac{1}{z_t} \sum_{i=1}^N \left( F_i(x_{t-1}) \right)^2 + \mu \right) \]

• Secondary image update:

\[ z_t = F^{-1} \left( \frac{1}{k_t} \sum_{i=1}^N \left( F_i(z_{t-1}) \right)^2 + \mu \right) \]

Results
• Quantitative evaluations on synthetic dataset

![Quantitative evaluations on synthetic dataset](image)

• Results on real-world examples (example patches cropped from result images)

• Results on spatially varying blur

• High computational efficiency

<table>
<thead>
<tr>
<th>Image size</th>
<th>256x256</th>
<th>512x512</th>
<th>1024x1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xu et al. (C+++)</td>
<td>14.8</td>
<td>34.4</td>
<td>271.9</td>
</tr>
<tr>
<td>Pan et al. (Matlab)</td>
<td>19.6</td>
<td>38.4</td>
<td>271.9</td>
</tr>
<tr>
<td>Hradil/Kels [2]</td>
<td>18.5</td>
<td>36.7</td>
<td>576.9</td>
</tr>
<tr>
<td>Ours (Matlab)</td>
<td>20.3</td>
<td>3.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Precomputation</td>
<td>1.8</td>
<td>4.6</td>
<td>15.3</td>
</tr>
</tbody>
</table>

• Robust to non-trained languages and severe page rotation

• Documents with color figures (benefit of estimating blur kernels)

1. Predict blur kernel of figure region from surrounding text regions;
2. Deblur figure region as a non-blind deconvolution problem.

Limitation and future work
• Our method cannot fully recover the details of an image degraded by large out-of-focus blur.

• Apply our method to other domain-specific images.

Reference
Multi-View Inverse Rendering under Arbitrary Illumination and Albedo

Kichang Kim    Akihiko Torii    Masatoshi Okutomi        Tokyo Institute of Technology

Introduction & the proposed method

Robust shape initialization by MVS
Refine it by solving the inverse rendering problem

Uniform albedo + common illumination
[Wu-CVPR11b]

Varying albedo + common illumination
[Zollhofer-ACMTOG16]

Varying albedo + per-image illumination
[Proposed]

< Pre-processing >
- Camera poses and initial shape estimation by SfM and MVS
  (VisualSFM [Wu-CVPR11a], CMPMVS [Jancosek-CVPR11])
- Surface subdivision (√3-subdivision [Kobbelt-ACMTOG00])
- Meshes are adaptively subdivided, such that all the projected areas of the triangular faces are smaller than a threshold.

< Joint estimation of illumination, albedo & shape >
- Minimize the cost

\[
\arg \min_{D, R, L} \left( \sum_{i} \left( \frac{|\mathbf{I}(i) - \mathbf{R}(i)|^2}{|\mathbf{I}(i)|} \right) + \alpha E_{\text{geom}}(D) + \beta E_{\text{geom}}(L) \right)
\]

- Geometric smoothness
  \[ E_{\text{geom}}(D) = \sum_{i} \left( \| \mathbf{K}(D, P(D, w_i)) \| \right) \]
- Smoothing the surface mesh by assuming locally planar surface where the weights are adjusted by the image gradient

- Photometric smoothness
  \[ E_{\text{geom}}(R) = \sum_{i} \left( \sum_{k \in \mathcal{N}(i)} (w_k R_{D}(i) - R_k)^2 \right) \]
- Smoothing the surface mesh by assuming the same albedo for neighboring vertices while detecting the borders of different materials:

- Chromaticity changes
  \[ w_{\text{chrom}} = -k_1 (1 - \sum_{k \in \mathcal{N}(i)} w_k) \]
- Intensity changes
  \[ w_{\text{int}} = -k_2 \sum_{k \in \mathcal{N}(i)} (R_{D}(i) - R_k)^2 \]

Experiments

The Joyful Yell (CG)

Input
Estimated albedo, illumination & shape
CMPMVS

The Yorkminster

Input
Estimated albedo, illumination & shape
CMPMVS

The Trevi Fountain

Input
Estimated albedo, illumination & shape
CMPMVS

Re-lighting using the estimated shape and albedo using the illumination estimated at different views

References
Deep Reconstruction-Classification Networks for Unsupervised Domain Adaptation

Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, Wen Li
mghifary@gmail.com, {bastiaan.kleijn, mengjie.zhang}@ecs.vuw.ac.nz, david.balduzzi@vuw.ac.nz, liwen@vision.ee.ethz.ch

European Conference on Computer Vision (ECCV2016)

1. Summary

✓ Propose a novel deep multi-task learning approach for unsupervised domain adaptation: Deep Reconstruction-Classification Networks (DRCN)
✓ DRCN jointly learns a shared encoding representation for two tasks: i) source output classification, ii) target input reconstruction, through backpropagation
✓ Achieve state-of-the-art performance on some cross-domain object classification tasks
✓ Reconstructed source images from DRCN resemble the appearance of target images
✓ Learning objective of DRCN is closely related to a semi-supervised learning framework, which leads to the soundness of the DRCN strategy

2. Related Work

Related work

✓ Unsupervised domain adaptation (UDA) is concerned with solving the dataset bias problem [TOK'11] in which labeled target data are not available
✓ Deep learning has played a major role in the advancement of domain adaptation in object recognition [DON'14, GAN'15], but we are still awaiting more powerful solutions

Background

✓ Generative model hypothesis: when labeled (target) data are lacking, modeling both the label and the structure of the data would induce a better discriminative model than modelling the label alone
✓ In UDA, however, labeled target data do not exist; we need to "borrow" labels from somewhere else for training
✓ Our method follows from the generative model hypothesis above by approximating target labels with source labels

3. Deep Reconstruction-Classification Networks (DRCN)

✓ Based on the aforementioned background, DRNC jointly learns two functions with a shared encoding representation:
  1. Labelling function \( f_{\theta_L} \) = \( (\theta_{L,s}, \theta_{L,t}) : X \rightarrow Y \) where the labeling function is conditioned on a labeled source domain and the reconstruction function is trained on an unlabeled target domain
  2. Reconstruction function \( f_{\theta_R} \) = \( (\theta_{R,s}, \theta_{R,t}) : X \rightarrow \hat{X} \)
✓ Given labeled source sample \( S^t = \{x^t, y^t\} \) and unlabeled target sample \( S^u = \{x^u\} \), define two empirical losses:
\[
\mathcal{L}(\theta_L) = \sum_{x^t} \ell(f_{\theta_L}(x^t), y^t) + \frac{1}{N_t} \sum_{x^u} D_{KL}(f_{\theta_L}(x^u) \parallel q_{\theta_L}(x^u))
\]

\( f_{\theta_L} \) and \( f_{\theta_R} \) are the classification loss and the reconstruction loss functions, respectively (e.g., cross-entropy loss and mean-squared loss), \( \theta_{L,s} \) and \( \theta_{L,t} \) are the model parameters.

✓ DRCN aims to solve the following objective
\[
\min \{\theta_L, \theta_R\} \mathcal{L}(\theta_L) + (1 - \lambda) \mathcal{L}(\theta_R)
\]

4. Algorithm

Learning algorithm: the network is trained by alternately minimizing \( \mathcal{L}(\theta_L) \) and \( \mathcal{L}(\theta_R) \) using backpropagation

5. Experiments

✓ 6 cross-domain object classification from some benchmarks: SVHN (SYN), MNIST (MN), USPS (US), STL-10 (ST), CIFAR-10 (CI)
✓ 8x8 performance gain over ReverseGrad [GAN'15] on SVHN and MNIST

6. Analysis

Relation to semi-supervised learning: Let \( p_D(x) \) be a parametric distribution with \( \theta \), DRCN can be interpreted probabilistically by assuming that \( p_D \) is Gaussian and \( q_{\theta_L}(x) \) is a multinomial distribution, fit by logistic regression. Thus, objective (1) is equivalent to the empirical MLE:
\[
\hat{\theta} = \arg \max_{\theta} \sum_{x^t} \log p_D(x^t | y^t) + \sum_{x^u} \log q_{\theta_L}(x^u | \hat{y})
\]

Recall a semi-supervised learning formulation proposed by [COC'06]: Suppose that samples are taken from the target distribution, i.e., \( x \sim q_T \), with probabilities \( \lambda \) taking the labeled samples and \( 1 - \lambda \) taking the unlabeled samples. Theorem 4.1 in [COC'06] states that semi-supervised learning solves the following MLE:
\[
\hat{\theta} = \arg \max_{\theta} \sum_{x^t} \log p_D(x^t | y^t) + (1 - \lambda) \sum_{x^u} \log q_{\theta_L}(x^u | \hat{y})
\]

We inspect a certain condition where (2) and (3) are closely related. Denote by \( A \) a source distribution with a covariate shift assumption \([H0'10]\): \( P_A(x) \not= P_B(x) \) are i.i.d. The empirical MLE (3) can be written as
\[
\hat{\theta} = \arg \max_{\theta} \sum_{x^t} \log p_D(x^t | y^t) + \sum_{x^u} \log q_{\theta_L}(x^u | \hat{y})
\]

The first term follows from the covariate shift assumption, while the second term defines an ergodic Markov chain whose asymptotic marginal distribution of \( X \) converges to \( P_B \).

Uselessness of unlabeled source samples: we argue that unlabeled source samples may not contribute to the cross-domain performance improvement, which explains that DRCN, EDAU, are no better than DRCN \( \text{ECAU} \) as a term of (4):
\[
\sum_{x^u} \log q_{\theta_L}(x^u | \hat{y}) \leq \sum_{x^t} \log p_D(x^t | y^t)
\]

On the second term, as \( \lambda \to 1 \), \( p_D(x) \) will converge to \( P_B \). Hence, since \( \sum_{x^u} \log q_{\theta_L}(x^u | \hat{y}) \) is adding more unlabeled source data will only result in a constant.

References


Architecture: a convolutional net as the labelling function and a convolutional autoencoder as the reconstruction function, where the encoding layers are shared.

In the experiment 1, the label prediction pipeline has three convolutional layers: 100 5x5 filters (CONV1), 150 5x5 filters (CONV2), and 200 3x3 filters (CONV3), two max-pooling layers of size 2x2 (POOL1 and POOL2), followed by three fully-connected layers (FC4, FC5, and FC_OUTPUT).
Identity Mappings in Deep Residual Networks
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
Microsoft Research Asia (MSRA)

Highlights
- Novel pre-activation residual structure
- Improved results using 1001-layer ResNet on CIFAR-10 and 200-layer on ImageNet

Importance of Identity Skip Connections
- ResNet with Identity mapping

Usage of Activation Function
- Post-activation to pre-activation

Analysis of Pre-activation Structure
- Ease of optimization
- Reducing overfitting

Code available:
- Deep Residual Networks with 1K Layers: https://github.com/KaimingHe/resnet-1k-layers

Results on CIFAR

<table>
<thead>
<tr>
<th></th>
<th>error (%)</th>
<th>CIFAR-10 error (%)</th>
<th>CIFAR-100 error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIN [9]</td>
<td>6.86</td>
<td>5.96</td>
<td>8.66</td>
</tr>
<tr>
<td>DNN [9]</td>
<td>6.62</td>
<td>5.88</td>
<td>8.54</td>
</tr>
<tr>
<td>PFA [10]</td>
<td>6.99</td>
<td>5.27</td>
<td>8.74</td>
</tr>
<tr>
<td>Highway [7]</td>
<td>7.56</td>
<td>5.54</td>
<td>8.84</td>
</tr>
<tr>
<td>ELU [16]</td>
<td>6.56</td>
<td>5.88</td>
<td>8.62</td>
</tr>
</tbody>
</table>

Results on ImageNet

<table>
<thead>
<tr>
<th></th>
<th>error (%)</th>
<th>top-1 (%)</th>
<th>top-5 (%)</th>
<th>top-1 (%)</th>
<th>top-5 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-101, original ResNet Unit</td>
<td>22%</td>
<td>27.7%</td>
<td>37.7%</td>
<td>22.0%</td>
<td>26.3%</td>
</tr>
<tr>
<td>ResNet-102, original ResNet Unit</td>
<td>19%</td>
<td>25.7%</td>
<td>35.4%</td>
<td>19.9%</td>
<td>24.3%</td>
</tr>
<tr>
<td>ResNet-103, proposed ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-104, original ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-105, proposed ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-106, original ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-107, proposed ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-108, original ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-109, proposed ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-110, original ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
<tr>
<td>ResNet-111, proposed ResNet Unit</td>
<td>18%</td>
<td>24.7%</td>
<td>33.4%</td>
<td>19.1%</td>
<td>23.6%</td>
</tr>
</tbody>
</table>

Analysis of Pre-activation Structure

<table>
<thead>
<tr>
<th>dataset</th>
<th>network</th>
<th>baseline unit</th>
<th>pre-activation unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-10</td>
<td>ResNet-101 (layer)</td>
<td>9.90</td>
<td>5.94</td>
</tr>
<tr>
<td></td>
<td>ResNet-110</td>
<td>6.61</td>
<td>5.97</td>
</tr>
<tr>
<td></td>
<td>ResNet-104</td>
<td>7.83</td>
<td>5.40</td>
</tr>
<tr>
<td></td>
<td>ResNet-1001</td>
<td>7.71</td>
<td>5.40</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>ResNet-101 (layer)</td>
<td>27.82</td>
<td>21.31</td>
</tr>
<tr>
<td></td>
<td>ResNet-104</td>
<td>25.10</td>
<td>21.31</td>
</tr>
</tbody>
</table>

Results on CIFAR

Results on ImageNet

Importance of Identity Skip Connections
- What if we break the identity shortcut?
- Various types of shortcut connections

Usage of Activation Function
- Post-activation to pre-activation

Analysis of Pre-activation Structure
- Ease of optimization
- Reducing overfitting

Code available:
- Deep Residual Networks with 1K Layers: https://github.com/KaimingHe/resnet-1k-layers

References
Deep Networks with Stochastic Depth
Gao Huang[1], Yu Sun[1], Zhuang Liu[2], Daniel Sedral[1], Kilian Weinberger[1]
*Equal contribution

Motivation
Training very deep networks is difficult:
• Gradients vanish and forward signals diminishes
• Long training time
• Overfitting

Question: Can we use short networks during training, but use deep networks during testing?

Idea: For each mini-batch, randomly drop a subset of layers and bypass them with the identity function!

Method
Stochastic depth network at training time
Mini-batch 1
Mini-batch 2
Mini-batch 3
A subset of layers are dropped at each mini-batch

$H_{\ell} = \text{ReLU}(b_{\ell} f_{\ell}(H_{\ell-1}) + \text{id}(H_{\ell-1}))$

• Linear decay rule for survival probabilities
  $b_{\ell} \sim \text{Bernoulli}(p_{\ell}) \quad \text{with} \quad p_{\ell} = (1 - \frac{\ell}{L}) \times 1 + \frac{\ell}{L} \times p_{L}$
  • Basic block (Similar to ResNets, He et al., CVPR’16)

Expected network depth
$E(\hat{L}) = \sum_{\ell=1}^{L} p_{\ell} = (3L - 1)/4 \approx 3L/4$

Stochastic depth network at test time
At test time

$H_{\ell}^{\text{Test}} = \text{ReLU}(b_{\ell} f_{\ell}(H_{\ell-1}^{\text{Test}}; W_{\ell}) + H_{\ell-1}^{\text{Test}})$

All layers are on, but outputs of $f_{\ell}$ are down weighted by their corresponding survival probabilities.

Advantages of stochastic depth
• Alleviates the gradient and signal vanishing problem
• Speeds up the training process
• Performs regularization and improves generalization (implicit ensemble of $2^L$ models)

Code
https://github.com/yueatsprograms/Stochastic_Depth

Classification
Training time

The gradient strength at the input layer

Hyper-parameter $p_{L}$
Varying $p_{L}$ with fixed depth
Varying $p_{L}$ with different depth

Analysis

Densely Connected Convolutional Networks (https://arxiv.org/abs/1608.06993)
• From implicit long-range connections to explicit long-range connections
• Learn more compact models!
• And more accurate!

Extension (DenseNets)

~25% faster

~25% shorter

DenseNets
Make the connections permanent!
**LESS IS MORE: TOWARDS COMPACT CNNS**

Hao Zhou\(^1\), Jose M. Alvarez\(^2\) and Fatih Porikli\(^3,4\)

\(^1\) University of Maryland, College Park, USA
\(^2\) Data61/CSIRO, Canberra, Australia
\(^3\) Australian National University, Canberra, Australia

**INTRODUCTION**

CNNs contain huge number of parameters, which leads to large memory footprint:

1. Fewer test samples at once.
2. Not suitable for Mobile devices.

Previous work:
1. Network distillation.
2. Memory efficient structures.
3. Parameter pruning.

We remove the number of neurons during training using sparse constraints. **removing neurons has advantages in**:

1. Do not rely on sparse data structure.
2. Also apply to Fourier domain.
3. Dimension reduction.

**CONTRIBUTIONS**

1. Reducing number of neurons of CNNs during training.
2. Analyzing the importance of ReLUs for sparse constraints.
3. Reducing significant amount of parameters for four well-known CNNs.
4. Easy to implement.

**SPARSE CONSTRAINTS**

Tensor Low Rank [2]:

\[
g(W) = \lambda \sum_{(i,j) \in \Omega} \frac{1}{n} \sum_{i=1}^{n} ||W_{ij}||_2, \quad \text{(3)}
\]

Group Sparsity:

\[
g(W) = \lambda \sum_{(i,j) \in \Omega} ||W_{ij}||_2, \quad \text{(4)}
\]

**IMPORTANCE OF ReLU**

Considering ReLU as:

\[
ReLU(x) = \begin{cases} 
\frac{x}{2} & \text{if } x > \epsilon \\
0 & \text{if } x \leq \epsilon.
\end{cases}
\]

then for a particular neuron \(\hat{W}_{ij} = 0\) is its local minimum if all other neurons are fixed.

**EXPERIMENTS**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conv2</th>
<th>Conv2</th>
<th>Conv3</th>
<th>Conv2</th>
<th>Conv3</th>
<th>Top-1 Error (%)</th>
<th>Parameter Reduction (%)</th>
<th>Memory Reduced (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet</td>
<td>60</td>
<td>45.5</td>
<td>97.5</td>
<td>60</td>
<td>56.5</td>
<td>97.5</td>
<td>0.54</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>46.5</td>
<td>97.5</td>
<td>80</td>
<td>56.5</td>
<td></td>
<td>0.77</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>47.5</td>
<td>97.5</td>
<td>100</td>
<td>56.5</td>
<td></td>
<td>0.76</td>
<td>0.03</td>
</tr>
<tr>
<td>CIFAR10</td>
<td>Quick</td>
<td>220</td>
<td>81.25</td>
<td>70.41</td>
<td>240</td>
<td>81.47</td>
<td>22.21</td>
<td>-0.12</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>240</td>
<td>81.47</td>
<td>70.31</td>
<td>240</td>
<td>81.47</td>
<td>22.73</td>
<td>0.4</td>
</tr>
<tr>
<td>VGG-13</td>
<td>40</td>
<td>30</td>
<td>56.49</td>
<td>50</td>
<td>30</td>
<td></td>
<td>44.58</td>
<td>-0.98</td>
</tr>
<tr>
<td></td>
<td>45</td>
<td>30</td>
<td>56.49</td>
<td>50</td>
<td>35</td>
<td></td>
<td>45.88</td>
<td>0.31</td>
</tr>
<tr>
<td>AlexNet</td>
<td>10</td>
<td>39.04</td>
<td>0.08</td>
<td>178.02</td>
<td>38.30</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>49.27</td>
<td>44.28</td>
<td>224.67</td>
<td>38.54</td>
<td>1.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>76.21</td>
<td>61.30</td>
<td>311.06</td>
<td>39.26</td>
<td>1.76</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results of LeNet on MNIST, CIFAR-10 quick on CIFAR-10 and AlexNet on ImageNet. Sparse constraints are added to two layers.

**COMPARE WITH [1]**

Right: compare with [1] on AlexNet.

[1] recursively combines similar neurons of a trained network.
When is Rotations Averaging Hard?
Kyle Wilson
David Bindel
Noah Snavely
kwilson24@washcoll.edu
bindel@cs.cornell.edu
snavely@cs.cornell.edu

The Rotations Averaging Problem

• A key subproblem in global Structure from Motion [2]

  Compute: vertex orientations $\mathcal{R} : V \to SO(3)$

  Given: a graph $G = (V, E)$ measures of relative orientation $\tilde{\mathcal{R}} : E \to SO(3)$

  To Minimize: squared geodesic error $\sum_{(i,j) \in E} d\left( R_{ij}, \tilde{R}_{ij}^T \right)^2$
  
  (this cost is very natural, but there are other common cost functions)

• Good solvers exist [3,4], but sometimes fail. Problem is nonconvex.

• Contributions: insights into which problems are easy, bounds on local convexity

Local Convexity

Solve failure $\leftrightarrow$ wrong local minima

Why can some problems be reliably solved, but others can’t? Which problems have bad local minima?

Not all nonconvex problems are equally hard.

Gauge Ambiguity

Old result: Rotations averaging is locally convex almost nowhere.

New result: Local convexity occurs in many problems of interest.

Key Idea: a gauge ambiguity reveals local convexity $(R_1, \ldots, R_n) \equiv (R_1S, \ldots, R_nS) \ \forall S \in SO(3)$

Toy Example:

$\min(1 - r)^2, 0 \leq r < 2, \theta \in [0, 2\pi)$

Note: locally non-convex on $0 \leq r < 1$

Fixing the Gauge: remove ambiguity by setting $\theta = 0$ (analogously, $R_k = 1$).

Results

Hessian of one term of the cost function:

$H_{ij} = \begin{bmatrix} \mu I + (2 - \mu)ww^T & -\mu I + (2 - \mu)ww^T \\ -\mu I + (2 - \mu)ww^T & \mu I + (2 - \mu)ww^T \end{bmatrix} + \begin{bmatrix} 0 & -\theta \bar{w} \bar{w}^T \\ 0 & 0 \end{bmatrix}$

1. positive semi-definite term (problem structure)
2. indefinite term (arises from curvature of the space)

Fixing the gauge makes the Hessian positive-definite in some parts of the problem domain.

Bounds:

Can we get insight by approximating away the directions of residuals?

$\lambda_{\min} \left( L_{\text{term}}^w \right) > 1 \implies \text{local convexity}$

where $L_{\text{term}}^w$ is a weighted, normalized graph Laplacian with it’s $x^0$ row and column removed.

And bound by the magnitude of the residuals too?

$\lambda_2(L)/\|w\| > \Delta/\mu_{\min} \implies \text{local convexity}$

Application

• Provides insight into good problem instance construction: high algebraic connectivity drives local convexity.

• Bounds are useful for predicting problem instance difficulty

References

CAPTURING DYNAMIC TEXTURED SURFACES OF MOVING TARGETS

ECCV'16

Ruizhe Wang¹  Lingyu Wei¹  Etienne Vouga²  Qixing Huang²,³  Duygu Ceylan⁴  Gerard Medioni¹  Hao Li¹

University of Southern California¹  University of Texas at Austin²  Toyota Technological Institute at Chicago³  Adobe Research⁴

ABSTRACT

We present an end-to-end system for reconstructing complete watertight and textured models of moving subjects such as clothed humans and animals, using only three or four handheld sensors. The heart of our framework is a new pair-wise registration algorithm that minimizes, using a particle swarm strategy, an alignment error metric based on mutual visibility and occlusion. We show that this algorithm reliably registers partial scans with as little as 15% overlap without requiring any initial correspondences, and outperforms alternative global registration algorithms. This registration algorithm allows us to reconstruct moving subjects from free-viewpoint video produced by consumer-grade sensors, without extensive sensor calibration, constrained capture volume, expensive arrays of cameras, or templates of the subject geometry.

SYSTEM OVERVIEW

VISIBILITY-BASED REGISTRATION

PRIOR WORK

template-based  costly capture setup  dynamic shape completion

REGISTRATION EVALUATION

SHAPE & TEXTURE RECONSTRUCTION

CAPTURE RESULTS
Location Recovery Problem:
Given: relative directions \( \{v_{ij}\} \) between cameras \( i, j \)
(for known camera orientations)
Find: camera locations \( \{t_i\} \)
Difficulty: many directions are outliers

Mathematical Formulation:
Let: \( t_1 \ldots t_n \in \mathbb{R}^3 \)
\( G = ([n], E = E_g \cup E_b) \)
\( v_{ij} = \frac{t_i - t_j}{\|t_i - t_j\|_2} \) for \( i \neq j \in E_g \)
\( v_{ij} \in S^2 \) for \( i, j \in E_b \)

Given: \( G, \{v_{ij}\} \)
Find: \( \{t_i\} \) up to translation and scale

Global SfM Pipelines:

Standard Pipeline:
- Estimate graph of Epipolar Geometries
- Solve for global camera orientations
- Solve for global camera locations
- Triangularize structure
- Bundle adjustment

Bipartite Pipeline:
- Estimate graph of Epipolar Geometries
- Solve for global camera orientations
- Set up a bipartite location recovery problem
- Solve for global camera and structure locations

ShapeFit:
A convex program for location recovery with outliers
minimize \( \sum_{ij \in E} \|P_{ij}(t_i - t_j)\|_2 \)
subject to \( \sum_{ij \in E} (t_i - t_j, v_{ij}) = 1, \sum_t t_i = 0 \)

ShapeKick:
A fast ADMM approach for ShapeFit using kicking
Augmented Lagrangian:
\( \mathcal{L}_\rho(T, Y, \lambda) = \sum_{ij \in E} \|P_{ij}(y_{ij})\|_2 + \frac{\rho}{2} \sum_{ij \in E} \|t_i - y_{ij} + \lambda_{ij}\|_2 \)
\( T \leftarrow \arg\min_{T \in G} \mathcal{L}_\rho(T, Y, \lambda), Y \leftarrow \arg\min_{Y \in \mathbb{R}^{(0,1)}} \mathcal{L}_\rho(T, Y, \lambda), \lambda_{ij} \leftarrow \lambda_{ij} + t_i - t_j - y_{ij} \)

ShapeFit: provably robust to outliers
Let: \( t_1 \ldots t_n \sim \mathcal{N}(0, I_{3 \times 3}) \) be i.i.d.
\( i, j \in E \) with prob. \( p = \Omega(n^{-1/5} \log^{3/5} n) \)
\( E_b \subset E \) be an arbitrary subset
\( v_{ij} \in S^2 \) be arbitrary for \( i, j \in E_b \)
\( \gamma = cp^5/\log^3 n \) for some \( c > 0 \)

Theorem
If \( n \) is large enough, and \( \max \deg(E_b) \leq \gamma n \),
then with probability at least \( 1 - \frac{1}{n^4} \),
the minimizer of Shapefit is unique and
exactly equals \( \{t_i\} \) up to translation and scale.

ShapeKick: comparable accuracy & 10x faster than state of the art

<table>
<thead>
<tr>
<th>Source</th>
<th>Median recovery error (m)</th>
<th>Solution time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC Library</td>
<td>2.2</td>
<td>1.2</td>
</tr>
<tr>
<td>Union Sq.</td>
<td>3.2</td>
<td>1.7</td>
</tr>
<tr>
<td>Montpelier</td>
<td>4.0</td>
<td>2.4</td>
</tr>
<tr>
<td>Tow. London</td>
<td>3.9</td>
<td>1.5</td>
</tr>
<tr>
<td>Notre Dame</td>
<td>3.5</td>
<td>3.3</td>
</tr>
<tr>
<td>Amaro</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Union Sq.</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Vienna Cdh.</td>
<td>4.3</td>
<td>7.6</td>
</tr>
<tr>
<td>Pisa</td>
<td>6.4</td>
<td>19</td>
</tr>
<tr>
<td>Piccadilly</td>
<td>1.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>

References:
Objectives

- We aim to develop a 3D shape representation by utilizing the heat flows on 3D surfaces and the corresponding temporal dynamics of the heat flows within the diffusion period.
- We employ LSTM to capture the temporal dynamics of heat flows and extract joint information between different time-steps that are either consecutive or with a large interval.
- We incorporate a 3-layer fully-connected neural network with LSTM, and extend the proposed approach to a cross-domain scenario.

Approach

The heat kernel $k_t(u,v)$ is introduced to measure the amount of heat that has been transformed from point $u$ to point $v$ on the 3D shape surface at time $t$, and it can be used to compute the HKS\[1\] of each vertex $u$ on the 3D shape surface at time $t$:

$$S_t(u,v) = k_t(u,u)$$

- HD-LSTM learns the heat kernel probability distribution over multiple diffusion time-steps, so as to capture the temporal dynamics.
- HD-LSTM minimizes the reconstruction between discriminative vectors $Y$ and the hidden unit outputs $h_t$ through time:

$$\arg\min_{W, U, V, b} \sum_{i=1}^{N} \sum_{t=1}^{T} ||Y - h_t||_2^2$$

Discriminative random vectors are generated from ground-truth shape labels, and are used to guide HD-LSTM toward learning discriminative 3D shape representations.

Source code: https://github.com/blacksmithfan/HD_LSTM

Experiments

McGill dataset (shape query)

<table>
<thead>
<tr>
<th>Methods</th>
<th>NN</th>
<th>1-Tier</th>
<th>2-Tier</th>
<th>DCG</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid BoW</td>
<td>0.95</td>
<td>0.63</td>
<td>0.79</td>
<td>0.88</td>
<td>-</td>
</tr>
<tr>
<td>Covariance method</td>
<td>0.97</td>
<td>0.73</td>
<td>0.81</td>
<td>0.93</td>
<td>-</td>
</tr>
<tr>
<td>Graph-based method</td>
<td>0.97</td>
<td>0.74</td>
<td>0.91</td>
<td>0.93</td>
<td>-</td>
</tr>
<tr>
<td>DeepShape [2]</td>
<td>0.98</td>
<td>0.78</td>
<td>0.83</td>
<td>0.70</td>
<td>0.46</td>
</tr>
<tr>
<td>BoW</td>
<td>0.90</td>
<td>0.41</td>
<td>0.54</td>
<td>0.90</td>
<td>-</td>
</tr>
<tr>
<td>HD-LSTM (without softmax)</td>
<td>0.97</td>
<td>0.88</td>
<td>0.83</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>HD-LSTM (with softmax)</td>
<td>0.08</td>
<td>0.92</td>
<td>0.95</td>
<td>0.95</td>
<td>0.04</td>
</tr>
</tbody>
</table>

SHREC14 dataset (sketch query)

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>FT</th>
<th>ST</th>
<th>DCG</th>
<th>AP</th>
</tr>
</thead>
</table>
| Source code: https://github.com/blacksmithfan/HD_LSTM

References

Multi-view 3D Models from Single Images with a Convolutional Network
Maxim Tatarchenko, Alexey Dosovitskiy and Thomas Brox
University of Freiburg

Overview

Our network learns a 3D representation of an object given a single input image of this object. It can predict an RGB image and a depth map of an object as seen from an arbitrary view. Multiple predicted views can be fused into a 3D point cloud.

Network architecture

Training data

The network was trained on renderings of synthetic 3D models.

Prediction results

3D model reconstruction

Multiple predicted views are fused into a single 3D point cloud, which can be further optimized to obtain a smooth mesh.

Natural input images

Even though trained on synthetic data, the network also generates reasonable predictions for natural input images.

Comparison with existing work

Our results compared to those from [1] (left), [2] (top right) and [3] (bottom right).

Interpolation in hidden space

Viewpoint dependency

More informative input views yield better predictions. Internal representation is insensitive to the input viewpoint change.

References

Linear Depth Estimation from an Uncalibrated, Monocular Polarisation Image

William A. P. Smith¹     Ravi Ramamoorthi²     Silvia Tozza³

¹ Department of Computer Science
University of York
² Center for Visual Computing
UC San Diego
³ Dipartimento di Matematica
Sapienza University of Rome

Overview

Overview of polarisation image processing.

Background

Shape-from-polarisation and polarisation images.

Degree of polarisation and shading cue

Degree of linear polarisation cue and shading cue.

Phase angle cue

We don't disambiguate locally. We write a linear equation that is satisfied by either interpretation.

Assumptions:
- Dielectric material (non-metallic)
- Orthographic projection
- Reflectance can be characterised as specular or diffuse dominant
- Known index of refraction (assumed = 1.5)
- Uniform albedo

Illumination estimation from a polarisation image

For the correct light source vector s:

\[ I_{\text{atm}} = \hat{\mathbf{t}}^T \hat{\mathbf{n}} \]

\[ \hat{\mathbf{n}} = \frac{\sin \theta \sin \phi}{\cos \theta} \]

Transformation between two possible polarisation normals

GBR transformation so solution ambiguous

Solving for depth

Finite difference approximation of surface gradient:

\[ p(x,y) = z(x+1,y) - z(x,y) \]

\[ q(x,y) = z(x,y+1) - z(x,y) \]

Stack all per-pixel equations in large linear system

\[ s^* = \arg \min \frac{1}{2} \left( \mathbf{Ax} - \mathbf{b} \right)^T \mathbf{Ax} - \mathbf{b} \]

Diffuse dominant pixel labelling

Specular pixels are phase shifted by \( \pi \)

Diffuse shading cue

Through the view of evidence:

\[ f_{\text{diffuse}} = \frac{\mathbf{w} \cdot \mathbf{h}}{\| \mathbf{w} \|^2 + \| \mathbf{h} \|^2 + 1} \]

First linear equation in surface gradient (diffuse):

\[ f_{\text{diffuse}} (\mathbf{w}, \mathbf{h}) = -p \mathbf{x} - q \mathbf{y} + s \]

First linear equation in surface gradient (specular):

\[ p = -\frac{h_x}{h_y}; \quad q = -\frac{h_y}{h_x} \]

Second linear equation in surface gradient:

\[ -p \cos \phi + q \sin \phi = 0 \]

Specular pixels are phase shifted by \( \pi \)

Azimuth aligns with \( \mathbf{w} \)

Experimental Results


Summary

Positives:
- Monocular
- Passive
- Uncalibrated illumination (can be spherical harmonic)
- Can be made single shot
- Depth directly by solving large, sparse linear system

Assumptions:
- Dielectric material (non-metallic)
- Orthographic projection
- Reflectance can be characterised as specular or diffuse dominant
- Known index of refraction (assumed = 1.5)
- Uniform albedo
Online Variational Bayesian Motion Averaging

Guillaume Bourmaud
Toshiba Research Europe

Introduction

Motion averaging (aka pose-graph inference for \( G = SE(3) \))

Given noisy relative transformations \( Z_{mk} \in G \)_{1:m<nk}

Estimate absolute transformations \( T_{iW} \in G \)_{i=1..N}

Example of application: RGB-D mapping

Visual Odometry

Loop closure detection

Online Motion Averaging

Aligned depth maps

Contributions

To perform online motion averaging on large scale problems, we propose an algorithm that is:

1. Computationally efficient: process the measurements one by one
2. Memory efficient: approximate the posterior distribution of the absolute transformations with a number of parameters that grows at most linearly over time
3. Robust: detect and remove wrong loop closures

Reparametrization of the absolute transformations

\[ T_{i(i+1)} = T_{iW} T_{i(i+1)W} \]

| Estimated transformations at time instant \( k \) | \( \{T_{iW}\}_{i=1..k} \) | \( \{T_{i(i+1)}\}_{i=1..k-1} \)
|---------------------------------|-----------------|-----------------
| Absolute                      |                 |                 |
| Relative                      |                 |                 |

Case of a single loop

\[
\arg\min_{(T_{iW})_{i=1..k}} \left\| \log G(Z_{iW} T_{iW} T_{i(i+1)W}^{-1}) \right\|_{2}^{2} + \sum_{j=i+1}^{N-1} \left\| \log G(Z_{jW} T_{jW} T_{j(i+1)W}^{-1}) \right\|_{2}^{2}
\]

odometry

loop closure

\[
\arg\min_{(T_{i(i+1)}_{i=1..k-1})} \left\| \log G(Z_{iW} \prod_{j=i+1}^{N-1} T_{jW} T_{j(i+1)W}^{-1}) \right\|_{2}^{2} + \sum_{j=i+1}^{N-1} \left\| \log G(Z_{jW} T_{jW} T_{j(i+1)W}^{-1}) \right\|_{2}^{2}
\]

odometry

loop closure

The relative parametrization induces very small correlations!

Motivation for a variational Bayesian approximation of the posterior distribution assuming independent relative transformation.

Online Variational Bayesian Motion Averaging Algorithm

- Approximated posterior at time \( k \rightarrow 1 \)
  \[ p(X_{k-1}|D_{o,k-1}, D_{c,k-1}) = \prod_{i=1}^{k} N_{\hat{T}_{i(i+1)}; \hat{T}_{i(i+1)}; P_{i(i+1)}} \]

- Processing of a new odometry measurement
  \[ p(X_{k}|D_{o,k}, D_{c,k-1}) = \prod_{i=1}^{k} N_{\hat{T}_{i(i+1)}; \hat{T}_{i(i+1)}; P_{i(i+1)}} \]
  where \( \hat{T}_{i(i+1)} = \hat{Z}_{o(i+1)} \) and \( P_{i(i+1)} = Z_{o(i+1)} \)

- Validation gating of a new loop closure measurement
  \[ p(Z_{a(i+1)}|D_{o,a(i+1)w}, D_{c,a(i+1)w}) = \prod_{i=1}^{k} N_{\hat{T}_{i(i+1)}; \hat{T}_{i(i+1)}; P_{i(i+1)}} \]
  \[ + \sum_{j=1}^{k} P_{i(i+1)} \]
  where \( \hat{T}_{i(i+1)} \) is obtained via a Gauss-Newton

- Processing of a new loop closure measurement
  \[ p(X_{a(i+1)}|D_{o,a(i+1)w}, D_{c,a(i+1)w}) = \prod_{i=1}^{k} N_{\hat{T}_{i(i+1)}; \hat{T}_{i(i+1)}; P_{i(i+1)}} \]
  \[ + \sum_{j=1}^{k} P_{i(i+1)} \]

Results

\[ G = SE(3): \text{Binocular 6D SLAM} \]

<table>
<thead>
<tr>
<th>RMSE position (m)</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>KITTI 00</td>
</tr>
<tr>
<td>Ours</td>
<td>2.1</td>
</tr>
<tr>
<td>COP-SLAM</td>
<td>6.0</td>
</tr>
<tr>
<td>LG-IEKF</td>
<td>0.8</td>
</tr>
<tr>
<td>g²拙</td>
<td>0.2</td>
</tr>
</tbody>
</table>

\[ G = S^3(3): \text{RGB-D Mapping} \]

\[ G = S\ell(3): \text{Video Mosaicing (see supplementary material)} \]

Ground Truth (Lidar)  Visual Odometry  COP-SLAM  Ours

Ours  Visual Odometry  COP-SLAM  Ours
Unified Depth Prediction and Intrinsic Image Decomposition from a Single Image via Joint Convolutional Neural Fields

Seungryong Kim¹, Kihong Park¹, Kwanghoon Sohn¹, and Stephen Lin²
¹Yonsei University, ²Microsoft Research

Introduction

- **Goal**
  - To jointly predict depth and intrinsic images from a single-image

  - Two tasks are formulated in a synergistic manner though CRF+CNN

Challenging

- **Depth Prediction**
  - No explicit depth cue
  - Depth ambiguity
  - No color-depth mapping assumption

- **Intrinsic Image Decomposition**
  - No prior
  - Retinex model
  - Occlusion
  - Shadows

Our Solution

- **Joint Depth**
  - Depth Prediction: Global Depth + Depth Gradient
  - Intrinsic Prediction: Albedo & Shading Gradient

- **Gradient Scale Network**
  - Confidence of Estimated Gradients
  - Color/Intrinsic Gradient → Depth Gradient Confidence

Formulation

- **Key-Insights**
  - Correlations are stronger among gradient domain

  - Value
    - Color
    - Depth
    - Albedo
    - Shading

  - Gradient
    - Color
    - Depth
    - Albedo
    - Shading

- **Network Configuration**
  - Depth Prediction Network
  - Gradient Scale Network

Unified Depth and Intrinsic Image Prediction

- **Training Procedure**
  - Global Depth Training
    \[ L(w_D^g) = \sum_{(i,p)} (D^g_p - F(i,p, w_D^g))^2 \]
  - Gradient Training + Gradient Scale Training
    \[ L(w_D^g, w_D^{gS}) = \sum_{(i,p)} \| D^g_p - G(i,p, w_D^{gS}) \|_2^2 \]
  - Iterative Training

- **Testing Procedure**
  - Iterative Joint Prediction
    \[ E(D) = \sum_p (D_p - D^g_p)^2 + \lambda G \sum_p \| D_p - C\nabla D^g_p \|_2^2 \]
    \[ E(A, S | I) = \sum_p (A_p - A^g_p - S_p)^2 + \sum_p \lambda_G \| A_p - C\nabla A^g_p \|_2^2 + \lambda_S \| S_p - C\nabla S^g_p \|_2^2 \]

Experimental Results and Discussion

- **Implementation Details**
  - VLFeat MatConvNet Toolbox

- **MPI SINTEL Benchmark**

- **NYU v2 RGB-D Benchmark**

- **Make3D RGB-D Benchmark**

Conclusion

- **Joint Convolutional Neural Field (JCNF) Model**
  - To jointly predicting a depth map and intrinsic images from single-image input

Contact

- E-mail: srkimg89@yonsei.ac.kr
- Homepage: http://diml.yonsei.ac.kr/~srkim/

European Conference on Computer Vision (ECCV) 2016
ObjectNet3D: A Large Scale Database for 3D Object Recognition
Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas and Silvio Savarese
Stanford University

- **Introduction**
  - Goal: build a large scale database for 3D object recognition
  - Application: recognizing the 3D properties of objects such as 3D location, 3D pose, 3D shape, etc.

- **Database Construction**
  - 100 rigid object categories
  - 2D image from the ImageNet database

- **Comparison with previous datasets**

- **Baseline Experiments**
  - Image-based 3D shape retrieval
  - Object proposal generation
  - Object detection and pose estimation

Acknowledgements. We acknowledge the support of NSF grants IIS-1528025 and DMS-1546206, and Google Focused Research Award, and grant SPO # 124316 and 1191689-1-UDAWF from the Stanford AI Lab-Toyota Center for Artificial Intelligence Research.
Branching Gaussian Processes
with Applications to Spatiotemporal Reconstruction of 3D Trees

Kyle Simek*, Ravishankar Palanivelu†, Kobus Barnard‡
*Matterport, Inc. †School of Plant Sciences, Univ. of Arizona ‡Computer Science, Univ. of Arizona

Abstract

Problem Statement
- Multi-view 3D reconstruction of plant structure
- Plants undergo significant nonrigid motion during imaging
- Cameras have position & orientation error

Model: Branching Gaussian Process
- Novel probabilistic model for branching tree structure
- Can model smooth curves under random motion
- Guaranteed attachment between curves
- Robust to small camera error

Approach
- Approximate Bayesian inference with expectation propagation
- Bayesian model selection to explore different topologies

Model

Representation: Curve tree
The curve tree function \( f(c,t) \) maps point index \((c,t)\) to 3D position.
Each point has an index \( (c,t) \) denoting curve index \( c \) and position \( t \).
The parent function, \( p(c,t) \) maps each point \((c,t)\) to its branch point, \((c',t')\).
In practice, we model densely sampled points \( k(c_1,t_1), (c_2,t_2), \ldots, (c_N,t_N) \)
with 3D positions \( z = \{ z_{c_1,t_1}, z_{c_2,t_2}, \ldots, z_{c_N,t_N} \} \).
To model motion, model one tree per time-step, \( t : \mathcal{Z} = \{ z_{i,t}, i \} \)
~11k dimensions (200-300 points / plant, 18 time points, 3D points)

Branching Gaussian Process Prior

\[
p(Z) = \mathcal{GP}(0,k(c,t,\tau,c',t',\tau')) \quad \text{where} \quad k(c,t,\tau,c',t',\tau') = k_x(c,t,c',t',\tau') + k_t(\tau,\tau')
\]
- Spatial covariance (recursive)
  \[
k_x(c,t,c',t',\tau') = \delta_{c,c'}k(t,t') + k_x(p(c,t),c',t',\tau')
\]
- Temporal covariance (Ornstein-Uhlenbeck process)
  \[
k_t(\tau,\tau') = \exp|\tau - \tau'|/\ell
\]
Properties:
1. Curves are \( c \) continuous.
2. Curves are attached at branch points.
3. Subtrees are independent, conditioned on branch point.

Training: Maximum marginal likelihood from single specimen.

Pixel Likelihood

Trained random forest classifier to extract foreground likelihood maps.
For each image:
1. Render tree \( z \) into image from time \( \tau \) as binary foreground image
2. Evaluate each binary pixel \( y_{ci} \) against likelihood map, independently.
   \[
p(D_i|z_{ci}) = \prod p(d_i|y_{ci})^{\gamma_{c_i}} p(d_i|y_{bg})^{1-\gamma_{c_i}}
\]
Full likelihood:
\[
L(Z) = p(Z|D_1:T_\tau) = \prod_{i=1}^{T} p(D_i|z_{ci})
\]

Inference

Goal: minimize \( p(Z|D_1:T_\tau) \times p(L(Z)) \)
- Intractable: \( L(Z) \) is non-convex, hundreds of dimensions.
- Approximate as block-diagonal Gaussian, \( L(Z) \) (3x3 blocks)
- Approximate posterior is \( q(Z) \propto p(Z)L(Z) \)
- Minimize KL divergence between \( q(Z) \) and \( p(Z|D) \)

Expectation propagation
1. Bootstrap initial guess for 3D model (see paper for details)
2. For each image \( D_i \):
   - Estimate point posterior \( q(z_i) \) by importance sampling
   - Extract Gaussian likelihood, construct full likelihood, \( \hat{L}(z_i) = \text{blockdiag}(L(z_i)) \)
3. Update \( q(Z) \) with new \( \hat{L}(z_i) \) (Kalman filter, RTS smoother)
4. Repeat 2, 3 until convergence
5. Extract \( p(D) \) for model selection

Model selection:
Propose birth/death of stems to optimize \( p(D) \) (see paper).

Results


References:

This work was supported by the Plant Collaborative under NSF Awards Numbers DBI-0735191 and DBI-1265383 and the Department of Education’s GAANN Fellowship.
Special thanks to Dr. Amy Tabb for sharing her shape from silhouette probability maps code for our evaluation.
Abstract

- Correlation filter (CF) trackers use circularly shifted patches as a proxy to real translations when training the filter.
- In all CF trackers, correlation scores are regressed to a Gaussian pulse centered around the previous location.
- This limits the trackers’ ability to recover from partial occlusion or to track fast moving objects.
- We propose a generic framework, in which we solve for both the optimal filter and target response based on exact translation detection scores from the image.
- The final formulation can use kernels and multiple templates jointly if solved in the dual domain. It improves all baseline CF trackers (SAMF, KCF, DCF, CSK, and MOSSE) by \(3 - 13\)% on OTB100 [1].

Motivation

In our formulation, we solve for both the filter \(w\) and the target response \(y\) in a unified objective. The target response is regressed to \(\hat{y}\). The resultant problem is solved in the dual domain with an efficient solution that allows the use of non-linear kernels and multiple templates jointly. The resulting optimization is:

\[
\begin{align*}
\min_{w, y} & \quad ||\hat{X}w - y||^2 + \lambda_1||w||^2 + \lambda_2||y - y_o||^2 \\
\text{s.t.} & \quad \hat{X} = \lambda_1E^{-1}G + \lambda_2E^{-1}G + \lambda_3E^{-1}G + \lambda_4E^{-1}G
\end{align*}
\]

The regressor \(y\) is generated out of a Gaussian interpolation to few samples generated from the correlation scores of exact translations in the image.

Problem Formulation

Experiments

We adopt our framework to five different baseline CF trackers: SAMF, KCF, DCF, CSK, and MOSSE. The performance of all baseline trackers improves significantly especially for the Fast Motion (FM), Motion Blur (MB), and Occlusion (OCC) categories. The experiments are done on the OTB100[1] dataset.

Learning Image Matching by Simply Watching Video

Gucan Long, Laurent Kneip, Jose M. Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu

Motivation

- Starting point: **Unsupervised learning of image matching**
- Applications: Feature matching, structure from motion, dense optical flow, recognition, motion segmentation, image alignment
- Problem: Difficult to do directly (e.g. based on video)
- Insights:
  - Image matching is a sub-problem of frame interpolation
  - Frame interpolation can be learned from natural video sequences

CNN: Unsupervised learning of frame interpolation

Lots of triplets of sequential frames

Input video

Training of CNN

Weak supervision signal: Temporal Coherence!

Architecture

Inspired by [FlowNet, Fischer et al.]

Results

Matching performance on KITTI flow training set

<table>
<thead>
<tr>
<th></th>
<th>MIND</th>
<th>DeepM</th>
<th>HoG</th>
<th>KLT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>APE</strong></td>
<td>5.838</td>
<td>3.240</td>
<td>7.856</td>
<td>8.836</td>
</tr>
<tr>
<td><strong>Accuracy@5</strong></td>
<td>0.719</td>
<td>0.875</td>
<td>0.688</td>
<td>0.808</td>
</tr>
<tr>
<td><strong>Accuracy@10</strong></td>
<td>0.876</td>
<td>0.951</td>
<td>0.875</td>
<td>0.864</td>
</tr>
<tr>
<td><strong>Accuracy@20</strong></td>
<td>0.948</td>
<td>0.977</td>
<td>0.947</td>
<td>0.906</td>
</tr>
<tr>
<td><strong>Accuracy@30</strong></td>
<td>0.967</td>
<td>0.986</td>
<td>0.964</td>
<td>0.927</td>
</tr>
</tbody>
</table>

Matching performance on MPI-Sintel training set

<table>
<thead>
<tr>
<th></th>
<th>MIND</th>
<th>DeepM</th>
<th>HoG</th>
<th>KLT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>APE</strong></td>
<td>4.695</td>
<td>3.442</td>
<td>9.680</td>
<td>8.157</td>
</tr>
<tr>
<td><strong>Accuracy@5</strong></td>
<td>0.716</td>
<td>0.835</td>
<td>0.455</td>
<td>0.702</td>
</tr>
<tr>
<td><strong>Accuracy@10</strong></td>
<td>0.915</td>
<td>0.953</td>
<td>0.805</td>
<td>0.826</td>
</tr>
<tr>
<td><strong>Accuracy@20</strong></td>
<td>0.981</td>
<td>0.987</td>
<td>0.929</td>
<td>0.903</td>
</tr>
<tr>
<td><strong>Accuracy@30</strong></td>
<td>0.993</td>
<td>0.993</td>
<td>0.959</td>
<td>0.938</td>
</tr>
</tbody>
</table>

Bold: Best result, underlined: Second best

MIND: our approach, DeepM: [DeepMatching, Weinzaepfel et al.]
A Distance for HMMs based on Aggregated Wasserstein Metric and State Registration

Yukun Chen, Jianbo Ye, and Jia Li
The Pennsylvania State University

1. Introduction

Hidden Markov Model

- M states: \{s_1, s_2, \ldots, s_M\}
- Transition matrix: \textbf{T}: \textbf{T}_{ij} = \mathbb{P}(s_{t+1} = j | s_t = i)
- Parameters of Gaussian probabilistic emission functions: \textbf{A}(\{\mu_i\}_{i=1}^{M}, \{\Sigma_i\}_{i=1}^{M}, \alpha)

We denote a HMM as:

\[ \Lambda(\textbf{T}, \textbf{M}) = \Lambda(\textbf{T}, \{\mu_i\}_{i=1}^{M}, \{\Sigma_i\}_{i=1}^{M}) \]

\(p\)-Wasserstein Distance

\[ W_p(f, g) = \inf_{\pi \in \Pi(f, g)} \left( \sum_{x \in X} d^p(x, y) \right)^{1/p} \]

If \( p = 2 \), \exists closed form solution for Gaussians:

\[ W_2(\phi, \phi) = \|\mu - \mu\|_2 + \text{tr}(\Sigma + \Sigma - 2(\Sigma \Sigma^T)^{1/2}) \]

2. State Registration

Treat as an optimal transport problem:

\[ \min_{\{w_i,j\}_{i,j=1}^{M}} \sum_{i,j=1}^{M} w_{i,j} W_2(\phi_i, \phi_j)^p \]

where

\[ \Pi(\pi_1, \pi_2) = \left\{ (W \in \mathbb{R}^{M \times M} : \sum_{i,j=1}^{M} w_{i,j} = \pi_{1,j}, j = 1, \ldots, M, \sum_{j=1}^{M} w_{i,j} = \pi_{i,1}, i = 1, \ldots, M) \right\} \]

3. MAW and IAW

1) Difference between GMMs:

\[ \tilde{R}_2(M_1, M_2; W)^p \triangleq \sum_{i,j=1}^{M} w_{i,j} W_2(\phi_i, \phi_j)^p \]

2) Difference between transition matrices:

\[ \tilde{T}_2^2 = \mathbb{W}_2^2(T_1, T_2)^p = \sum_{i,j=1}^{M} \tilde{w}_{i,j} W_2(\phi_i, \phi_j)^p \]

Aggregated Wasserstein:(MAW/IAW)

\[
\tilde{R}_2(A_1, A_2; W)^p = (1 - \alpha)\tilde{R}_2(M_1, M_2; W) + \alpha D_2(T_1, T_2; W)
\]

4. Experiments

4.1 Synthetic Data: (1-NN Retrieval)

Table: Summary of the parameters setup for parameter perturbation experiments. \( \text{rand}(\mathbb{Z}) \) here means generating random samples from Dirichlet distribution with parameter \( \alpha \).

<table>
<thead>
<tr>
<th>( \delta )</th>
<th>( \mu )</th>
<th>( \Sigma )</th>
<th>( \alpha )</th>
<th>( T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

5. References

- [Jiang et al.](#)
- [Lv et al. ECCV 2006](#)

---

Figure: Visualization of CMU motion capture data. Left: Jump. Right: Walk

Figure: Precision-recall plot for Motion Retrieval. The plot for 6 joint-groups, i.e., \( \text{root}, \text{head}, \text{neck}, \text{thorax}, \text{elbow}, \text{body}, \text{leg}, \text{leg} \), are displayed separately.

---

Figure: Testing accuracies w.r.t iteration number of Adaboost (number of weak classifiers selected). (a) Motion Classification by Adaboost on 6 joints. (b) Motion Classification by Adaboost on 27 joints. The iteration number means the number of features incrementally acquired in Adaboost.

---

Table: Summary of the parameters setup for parameter perturbation experiments. \( \text{rand}(\mathbb{Z}) \) here means generating random samples from Dirichlet distribution with parameter \( \alpha \).

<table>
<thead>
<tr>
<th>( \delta )</th>
<th>( \mu )</th>
<th>( \Sigma )</th>
<th>( \alpha )</th>
<th>( T )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

---

Figure: (a) Experiment scheme for varying \( \mu \) and varying \( \Sigma \). A re-estimated \( \phi \) is denoted as the dashed blue line. (b) and (c) Mean estimates of \( W_2(\phi_0, \phi_1) \) (blue) and \( KLD(\phi_0, \phi_1) \) (orange) and their 1σ confidence intervals w.r.t different Gaussian \( \phi_0 \). (b) is for varying \( \mu \), and (c) is for varying \( \Sigma \).
Motivation & Contributions

• We propose an integrated, fully-differentiable deep network, for keypoint detection, orientation estimation and feature description.
• Joint optimization improves overall performance.
• Outperforms the state-of-the-art on multiple datasets.
• Provides an off-the-shelf replacement for SIFT, with a practical computational time: 1.5x-3x that of SIFT.
• Code is available: https://github.com/cvlab-epfl/LIFT

Training with Patches

• Train with patches to make the problem tractable and scalable.
• Two SfM datasets: Piccadilly ('pic') and Roman Forum ('rf').
• Keep only SfM points, i.e. we learn to find repeatable points.

Quadruplet Siamese Network

• Training patches on SIFT locations, perturbed to avoid biases.
• Quadruplet: training the full pipeline requires non-keypoints, matching keypoints, and non-matching keypoints.

A Single Cost Function

\[
\begin{align*}
\min_{f_p, g_p, h_p} & \quad \mathcal{L}_{\text{class}}(P^1, P^2, P^3, P^4) + \gamma \mathcal{L}_{\text{pairwise}}(P^1, P^2, P^3) \\
\text{detector} & \text{orientation} \text{ descriptor} \\
\sum_{i=1}^4 & \max(0, 1 - \text{softmax}(f_p(P^i))) y_i)^2 \\
\text{descriptor} & \text{detector} \text{ descriptor} \\
& \left\| h_p(G(P^1, \text{softmax}(f_p(P^i)))) - h_p(G(P^2, \text{softmax}(f_p(P^i)))) \right\|_2^2 \\
& + \max(0, C - \left\| h_p(G(P^1, \text{softmax}(f_p(P^i)))) - h_p(G(P^3, \text{softmax}(f_p(P^i)))) \right\|_2) \\
& G(P, x) = \text{Rot}(P, x, g_p(\text{Crop}(P, x)))
\end{align*}
\]

Spatial Transformers (Rot/Crop) are used as differentiable tools for image transformations. Note that these modules are not trained.

Integrated LIFT Network

• DET, ORI, DESC: Based on state-of-the-art deep networks.
• Differentiable “Glue”: Spatial Transformers & softargmax.

Run-time Pipeline

Detector runs in scale-space with Non-Maximum Suppression. The Orientation Estimator and Descriptor only process keypoints.

Joint Optimization

• Descriptor performance, in terms of NN mAP.
• LIFT descriptor works best with LIFT keypoints.
• Joint optimization is key.

Evaluation

• ‘Strecha’: wide-baseline stereo (urban scenes).
• ‘DTU’: viewpoint changes (objects).
• ‘Webcam’: natural illumination changes, same viewpoint (outdoor).
• Metric: Matching score to capture full-pipeline performance.
• The ratio of correct matches recovered in the shared viewpoint region.
• Results: best performance on all datasets, with ‘rf’ and ‘pic’.

LIFT: Learned Invariant Feature Transform

Kwang Moo Yi*,1), Eduard Trulls*,1), Vincent Lepetit2), and Pascal Fua1)
1) École Polytechnique Fédérale de Lausanne 2) Graz University of Technology (* Equal contribution)
kwang.yi@epfl.ch, eduard.trulls@epfl.ch, lepetit@icg.tugraz.at, pascal.fua@epfl.ch
Learning a Predictable and Generative Representation for Objects
Rohit Girdhar, David Fouhey, Mikel Rodriguez and Abhinav Gupta

Motivation
What is the right object representation?

Good object representation should satisfy both:
• Generative in 3D
• Predictable from 2D

T-L Embedding Architecture

Reconstruction Results

Quantitative Results
Reconstruction evaluation using Average Precision over voxel prediction

Latent Space Analysis
Interpolating between two random models generates plausible structures
Generating new models by composing two latent representations

Fast 3D Model Retrieval
Nearest neighbor over latent representation for 3D models and images
General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues

Helge Rhodin\textsuperscript{1} Nadia Robertini\textsuperscript{1,2} Dan Casas\textsuperscript{1} Christian Richardt\textsuperscript{1,2,3} Hans-Peter Seidel\textsuperscript{1} Christian Theobalt\textsuperscript{1}
\textsuperscript{1}Max Planck Institute for Informatics  \textsuperscript{2}Intel Visual Computing Institute \textsuperscript{3}Now at University of Bath

Contributions

- Fully automatic actor shape and pose estimation in general scenes, unsupervised motion
- Joint parametric model
  - Volumetric density, mesh, and skeleton
- Spatio-temporal optimization
  - Generative & discriminative, top-down & bottom-up

Challenges

- Unknown actor
- Unknown motion
- Diverse body shapes
- Dynamic background
- Color ambiguities
- Few views

Outdoor reconstructions

- Input views
- Colored density
- Actor skeleton and mesh

Indoor reconstructions

- Real
- Volume
- Skeleton
- Mesh

Volumetric contour cues (Stage II)

- Stage I: Multi-view input
- Stage II: Image gradients

Evaluation - Shape

Silhouette overlap metric (Human Eva)

Body dimension metric

Evaluation - Pose

Human Eva (3D joint position error in mm)

Two-view reconstruction

Related work


Funding: ERC starting grant CapReal (335545)
Globally Continuous and Non-Markovian Crowd Activity Analysis from Videos

He Wang & Carol O’Sullivan
h.e.wang@leeds.ac.uk & Carol.O’Sullivan@scss.tcd.ie
Disney Research LA, US & University of Leeds, UK & Trinity College Dublin, Ireland

Abstract

Automatically recognizing activities in video is a classic problem in vision and helps to understand behaviours, describe scenes and detect anomalies. We propose an unsupervised method for such purposes. Given video data, we discover recurring activity patterns that appear, peak, wane and disappear over time. By using non-parametric Bayesian methods, we learn coupled spatial and temporal patterns with minimum prior knowledge. To model the temporal changes of patterns, previous works compute Markovian progressions or locally continuous motifs whereas we model time in a globally continuous and non-Markovian way. Visually, the patterns depict flows of major activities. Temporally, each pattern has its own unique appearance-disappearance cycle. To compute compact pattern representations, we also propose a hybrid sampling method. By combining these patterns with detailed environment information, we interpret the semantics of activities and report anomalies. Also, our method fits the data better and detects anomalies that were difficult to detect previously.

Introduction

Understanding crowd activities from videos has been a goal in many areas. The main problem is essentially mining recurrent patterns over time from video data, shown in Fig. 1.

Methodology

We compute activities as flows consists of bundles of similar trajectories. To feed trajectories into our topic model, we first convert video data into text data. By discretizing the space and velocity domain temporally it is not an anomaly but temporally it is, which is difficult to detect for existing methods.

Experimental Results

We tested our method on three widely used datasets: Forum [2], Carpark [3] and TrainStation [4]:

Activity Computation

Comparison

By comparison, our method provides better likelihoods (Tab. 1 Left) and more in alignment with human judgements (Tab. 1 Right). r_correct and r_complete are computed against human labelling.

Anomaly Detection

Figure 4: Top Left: Environment of the car park. Bottom Left: Observation numbers over time Right: Some activities shown by representative trajectories and their respective time activities. Colors indicate orientations described by the legend in the middle.

Figure 5: Top Left: Environment of the New York Central Terminal. Bottom Left: Observation numbers over time Right: Some activities shown by representative trajectories and their respective time activities. Colors indicate orientations described by the legend on the right.

Anomaly Not Detected By Existing Methods

Figure 7: Left: A pattern-detected. Middle: The temporal profile of the pattern. Right: An anomaly detected. Note spatially it is not an anomaly but temporally it is, which is difficult to detect for existing methods.

References

1. Introduction

- Existing studies tackle the problems of face alignment and 3D face reconstruction separately.
  - Face alignment methods mostly do not consider the visibility of landmarks. Consequently, their performance degrades for large pose face images.
  - 3D face reconstruction methods suffer from the invisible landmarks, and do not generate normalized 3D face for face recognition purpose.

2. Proposed Algorithm

- The coupled process in each iteration
  - Step1: Updating landmarks. The adjustment to the landmarks’ locations is determined by the local texture features via a landmark regressor.
  - Step2: Updating 3D shape. The above-obtained landmark location adjustment is used to estimate the adjustment of the 3D shape via a shape regressor.
  - Step3: Refining landmarks. Once the 3D shape is updated, the landmarks with visibility can be further refined with a 3D-to-2D mapping matrix.

- Cascaded coupled-regressor learning
  - Landmark regressors: We employ linear regressors as the landmark regressors:
    \[
    R^l_i = \arg \min_{R^l_i} \sum_{k \in K} \left\| U^l_i - U^l_k \right\|^2 - R^l_i (kU^l_i, U^l_k)
    \]
  - Shape regressors: Using similar linear regressors in shape space [1]:
    \[
    R^s_i = \arg \min_{R^s_i} \sum_{k \in K} \left\| S^s_i - S^s_k \right\|^2 - R^s_i (M^s_i)
    \]
  - 3D-to-2D mapping and landmark visibility
    - We assume that the expression and pose induced deformation can be approximated by a linear transform. The visibility of landmarks are computed based on the surface normal and the camera rotation matrix [2].

3. Training Data

- Two different models are trained using two training sets.
  - BU3DFE #100 subjects
    - 2D images: #13,300
      (19 poses, 7 expressions)
    - 3D faces: #100
      (frontal, neutral)
  - LFW #150 subjects
    - 2D images: #4,149
    - 3D faces: #150
      (frontal, neutral)

- 3D shapes in BU3DFE have been established dense correspondence by [3].
- The 68 2D landmarks of LFW images are provided by the work of [4].
- The 3D neutral shapes of LFW are obtained by [1].

4. Experimental Results

- Face alignment accuracy on AFW dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>NME</td>
<td>7.52%</td>
<td>5.60%</td>
<td>3.15%</td>
</tr>
</tbody>
</table>

- 3D face reconstruction accuracy on BU3DFE dataset

- Application to face recognition on BU3DFE dataset

- The effect of the reconstructed PEN 3D face shapes on face recognition is evaluated by performing direct 3D-to-3D matching and fusion with conventional 2D face recognition.
- Computational efficiency
- The Matlab implementation on a PC with i7-4710 CPU runs at ~26 FPS.

5. Summary

- By alternately applying cascaded landmark regressors and 3D shape regressors, the proposed method can effectively accomplish the two tasks of face alignment and 3D face reconstruction simultaneously in real time.
- Extensive experiments with comparison to state-of-the-art methods demonstrate the effectiveness of the proposed method in both face alignment and 3D face shape reconstruction, and in facilitating face recognition as well.

References:
Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image

Federica Bogo*,2, Angjoo Kanazawa*,3, Christoph Lassner1,4, Peter Gehler1,4, Javier Romero1, Michael Black1

MPI for Intelligent Systems1, Microsoft2, University of Maryland3, University of Tübingen4

* - Equal contribution.
The work was performed at the MPI for Intelligent Systems.

Acknowledgements: We thank M. Al Borno for inspiring the capsule representation, N. Mahmood for help with the figures, I. Akhter for helpful discussions.

References:
5. Ramakrishna, V., Kanade, T., Eshghi, Y.: Reconstructing 3D Human Pose from 2D Image Landmarks. ECCV 2012

Problem
Fully automatic 3D pose and shape estimation from a single image

Challenges:
• Complex poses in natural images
• Monocular
• Accuracy required

Past work:
• Manual intervention [1]
• 3D stick figures [4,5,6]

We estimate 3D pose & shape, as a full 3D mesh by:
1. 2D joint estimation via CNN [2]
2. 3D generative model SMPL [3]

Intuition: What makes it work?
SMPL shape + pose prior + interpenetration reasoning

Fit from just 2D joints

SMPL Body Model [3]
Joints are regressed from pose & shape

shape: 10 PCA coefficients
pose: 23 joint angles

Interpenetration Reasoning
Body ≈ 20 capsules (axis length + radius)
• Train a regressor: \( \beta \rightarrow \) capsules
• Capsules are posed according to \( \theta \)
  • differentiable wrt to \( \beta \) and \( \theta \)
• Penalize capsule intersection

Objective Function

\[
E(\beta, \theta) = E_J(\beta, \theta; K, J_{est}) + E_\alpha(\theta) + E_{\theta}(\theta) + E_{\theta}(\theta; \beta) + E_\beta(\beta)
\]

Step 1: Set up Camera \( K = \{ R, t \} \)
• Perspective camera with known / approx. focal length
• Optimize \( E_J \) on torso joints wrt \( K \)

Step 2: Solve for Pose & Shape
• Minimize \( E(\beta, \theta) \)

Evaluation

Synthetic: Shape
Data: Projected joints from 1000 synthetic 3D models + noise.
Metric: Mean vertex-to-vertex L2 distance between the estimated and true shape in a canonical pose.

Quantitative: Pose
Data: HumanEva & Human3.6M
Metric:
• Mean L2 distance between ground truth and predicted 3D joints after preprocessing analysis
• Compared with 3 state-of-the-art methods [4,5,6].
• CNN was not fine-tuned for these datasets
• All methods use the same detected 2D joints as input

Qualitative: Leeds Sports Pose (LSP)

Failure Cases

Code & Results @ smplify.is.tue.mpg.de

What’s next?
Silhouettes, Multiple Views, Video
Multiple people, Occlusion Reasoning, Data Generation

Challenges:
• Complex poses in natural images
• Monocular
• Accuracy required

SMPLify

Our approach:

Joints are regressed from pose & shape

Input
Before
After

• 2D joint estimation via CNN [2]
• 3D generative model SMPL [3]
Do We Really Need to Collect Millions of Faces for Effective Face Recognition?

Iacopo Masi*,1, Anh Tuan Tran*,1, Tal Hassner*2,3, Jatuporn Toy Leksut1 and Gerard Medioni1

1. Institute for Robotics and Intelligent Systems, USC, CA, USA
2. Information Sciences Institute, USC, CA, USA
3. The Open University of Israel, Israel

* Denotes equal authorship

A. The two keys to effective face recognition

1. During training: Learn the variability of same-subject appearances

Increase training set intra-subject appearance variations

2. During testing: Make same subjects easier to compare

Reduce test set intra-subject appearance variations

B. The problem with existing training sets

<table>
<thead>
<tr>
<th>Method</th>
<th># train img.</th>
<th># subj.</th>
<th>#img. / subj.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFaces'14 (FB)</td>
<td>4 m</td>
<td>4,030</td>
<td>1k</td>
</tr>
<tr>
<td>VGG Face'15</td>
<td>2.6 m</td>
<td>2,622</td>
<td>1k</td>
</tr>
<tr>
<td>Face Net’15 (Google)</td>
<td>200 m</td>
<td>8 m</td>
<td>25</td>
</tr>
<tr>
<td>Fusion'15 (FB)</td>
<td>500 m</td>
<td>10 m</td>
<td>50</td>
</tr>
<tr>
<td>MegaFace'16</td>
<td>1.02 m</td>
<td>690.5k</td>
<td>15</td>
</tr>
</tbody>
</table>

Even with lots of resources, it’s hard to ensure sufficient intra-subject and pose variations.

C. Domain (face) specific data augmentation

For each CASIA image, synthesize 3 types of new images without changing the subject label:

I. Pose variations

II. 3D shape variations

III. Expression variations

The new augmented set

Far more inter-subject appearance variability!

D. Matching by reducing test set intra-subject appearance variability

E. Results

 effect of augmentation on training

Training better CNNs with less effort using domain (face) specific data augmentation!!!

For code, data, more results and info, see https://goo.gl/RYm3xU

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA 2014-14071600010.
Poster Sessions
Fundamental Matrices from Moving Objects Using Line Motion Barcodes

Yoni Kasten  Gil Ben Artzi  Shmuel Peleg  Michael Werman
The Hebrew University of Jerusalem

**Introduction**

The Goal: Calibration of multi-camera systems from significantly different viewpoints, when the scene has multiple moving objects.

Prior Work: When finding corresponding points is difficult, use corresponding epipolar lines from dynamic silhouettes. Motion Barcode are used in order to accelerate the search.

Our Contribution: When there are multiple objects in the scene, previous methods fail. Our method can handle such cases by sampling lines.

**Prior Work**

Ben-Artzi et al. Camera Calibration from Dynamic Silhouettes Using Motion Barcodes, CVPR'16.

- Uniformly sampling tangent lines around the convex hull of the silhouette in each frame.
- In each frame time, tangent lines from one camera view are compared to tangent lines from second camera view.
- The pair of lines having highest motion barcode similarity is selected.
- Number of RANSAC iterations is much lower than previous papers (Sinha & Pollefeys, IJCV 2010).
- Limitation: With multiple objects in the scene – convex hulls do not include the same objects.

**Motion Barcodes for Lines (Ben-Artzi et al)**

For a segmented video (moving objects vs. static background), For line \( \ell \) at frame \( t \):

\[
\begin{align*}
\text{if} & \quad \ell \text{ intersects object} \\
\text{else} & \quad 0
\end{align*}
\]

Barcode correlation defines the temporal similarity between two lines:

\[
d_{\ell_1, \ell_2}(t) = \text{corr}(b_\ell(t), b_\ell(t'))
\]

Corresponding epipolar lines have correlated Motion Barcodes.

**Method**

- The Goal: Find the Fundamental Matrix between two cameras viewing a dynamic scene.
- Lines are sampled in both cameras by connecting two uniformly sampled points on the border of the frame (see picture).
- Motion Barcodes are computed for all sampled lines.
- Only informative lines are used in next steps, i.e. lines whose motion barcodes have enough zeros and ones. This leaves \( n_1 \) line Motion Barcodes for Camera A, and \( n_2 \) line Motion Barcodes for Camera B.
- The correlations of all line motion barcodes from Camera A with all those from Camera B are computed:

\[
\text{correlation matrix of size } n_1 \times n_2
\]

- If the correlation of a pair of lines is in the mutual top 3 of each other, i.e. top 3 in both row and column, it is considered a candidate.
- The 1,000 candidate pairs having the highest correlations are taken as candidates for corresponding epipolar lines.
- As a result we get 1000 pairs of corresponding epipolar line candidates: \( \{(i_1, i'_1), ..., (i_{1000}, i'_{1000})\} \)
- Experimentally the probability that a candidate pair is true, i.e. both lines are correct epipolar lines, is 0.7 for synthetic data, and 0.37 for real data.

**RANSAC samples**

In each RANSAC iteration:

- two candidate pairs of lines are randomly sampled: \( (i_1, i'_1) \) and \( (i_2, i'_2) \)
- The intersection of each two lines in each picture gives two candidate epipoles:

\[
e = i_1 \times i_2, \quad e' = i'_1 \times i'_2
\]

- Another pair of lines passing \( (i_3, i'_3) = \arg \min_{(i, i') \in \text{candidates}} d(l_i(e), d(l_i'(e')) \)

through the epipoles is chosen:

- The epipolar line homography \( H \) is calculated using the DLT algorithm, and its consistency with all other candidates is measured

For a candidate pair \( (i, i') \): The image area between \( H_i \) and \( i' \) is used as a similarity measure.

The pair \( (i, i') \) is considered an inlier if the area is small enough.

The Fundamental Matrix \( F \) is computed from the most consistent epipolar line homography:

\[
F = [e']_2 H^{-T}
\]

**Results**

- Running 10,000 RANSAC iterations for each pair of cameras in each dataset resulted in a Fundamental Matrix for each camera pair.
- The Symmetric Epipolar error of the resulting \( F \) was measured on ground truth points.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Average error</th>
<th>Number of good pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cubes</td>
<td>0.31</td>
<td>10/10</td>
</tr>
<tr>
<td>Thin Cubes</td>
<td>0.79</td>
<td>21/21</td>
</tr>
<tr>
<td>Pets 2009</td>
<td>1.09</td>
<td>5/6</td>
</tr>
</tbody>
</table>
Human Pose Estimation using Deep Consensus Voting

Ita Lifshitz, Ethan Fetaya and Shimon Ullman

Overview:

- Each image location votes for the position of all keypoints using a deep CNN.
- The voting scheme utilizes information from the whole image, rather than rely on a sparse set of locations.
- We compute image-dependent joint keypoint probabilities by looking at consensus voting.
- Inference is performed sequentially, starting from a subset of reliable keypoints and finishing in full inference.

1. The voting scheme

- We discretize the votes into log-polar bins around the patch center $y$.
- Our CNN outputs a distribution over log-polar bins, for each keypoint.
- Each vote gives a rough estimate for the keypoint location.
- The aggregated votes give precise estimation for keypoint location.
- We use GPU deconvolution implementation to efficiently aggregate the votes to a probability distribution

$$P(K_j = x) = \sum_{y \in Y} P_y(K_j = x) = \text{deconv}(s^j, w)$$

2. Image-dependent binary probabilities

- At each location $y$, the votes for all keypoints are assumed to be independent.

$$P_y(K_i = x_i, K_j = x_j) = P_y(K_i = x_i) \cdot P_y(K_j = x_j)$$

- After aggregating the votes from all locations $y$, the joint probabilities become dependent through the voters.

$$P(K_i = x_i, K_j = x_j) \propto \sum_y P_y(K_i = x_i) \cdot P_y(K_j = x_j)$$

- We call this the consensus voting term since for joint probability to be high, the combination needs to get strong votes from many common voters.

3. Inference

- We minimize the total energy function $\sum_{i=1}^{N} \phi_i(x_i) + \sum_{(i,j) \in F} \phi_{ij}(x_i, x_j)$ using TRW-S
- Local geometric constraint of a middle point between keypoints.
- Sequential prediction of keypoints: starting head to pelvis, then adding the torso and finally the whole pose.

4. Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tonso e et al.</td>
<td>95.8</td>
<td>90.3</td>
<td>90.5</td>
<td>94.3</td>
<td>77.6</td>
<td>60.7</td>
<td>62.8</td>
<td>79.6</td>
</tr>
<tr>
<td>Carreira e et al.</td>
<td>95.1</td>
<td>91.7</td>
<td>91.7</td>
<td>72.4</td>
<td>82.8</td>
<td>74.2</td>
<td>66.6</td>
<td>89.6</td>
</tr>
<tr>
<td>Pishchulin e et al.</td>
<td>94.1</td>
<td>90.2</td>
<td>93.4</td>
<td>77.3</td>
<td>82.6</td>
<td>75.7</td>
<td>86.8</td>
<td>82.1</td>
</tr>
<tr>
<td>Wei e et al.</td>
<td>97.8</td>
<td>95.0</td>
<td>98.7</td>
<td>98.6</td>
<td>98.8</td>
<td>97.4</td>
<td>98.6</td>
<td>98.6</td>
</tr>
<tr>
<td>Nsnewe e et al.</td>
<td>97.6</td>
<td>95.4</td>
<td>96.0</td>
<td>95.2</td>
<td>98.7</td>
<td>85.0</td>
<td>80.0</td>
<td>89.4</td>
</tr>
<tr>
<td>Our Model</td>
<td>97.8</td>
<td>93.3</td>
<td>95.7</td>
<td>90.4</td>
<td>85.3</td>
<td>76.6</td>
<td>70.2</td>
<td>55.0</td>
</tr>
</tbody>
</table>

MPII Human Pose Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Torso</th>
<th>U. Leg</th>
<th>L. Leg</th>
<th>U. Arm</th>
<th>Forearm</th>
<th>Head</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tonso e et al.</td>
<td>95.3</td>
<td>91.4</td>
<td>90.1</td>
<td>93.0</td>
<td>95.2</td>
<td>93.7</td>
<td>88.6</td>
</tr>
<tr>
<td>Carreira e et al.</td>
<td>95.3</td>
<td>81.8</td>
<td>73.3</td>
<td>66.7</td>
<td>51.0</td>
<td>84.1</td>
<td>72.5</td>
</tr>
<tr>
<td>Cheh Y of e et al.</td>
<td>96.0</td>
<td>77.2</td>
<td>72.2</td>
<td>99.7</td>
<td>58.1</td>
<td>85.6</td>
<td>83.6</td>
</tr>
<tr>
<td>Fan e. et al.</td>
<td>95.4</td>
<td>77.7</td>
<td>90.8</td>
<td>92.8</td>
<td>90.1</td>
<td>90.1</td>
<td>96.0</td>
</tr>
<tr>
<td>Pishchulin e et al.</td>
<td>97.0</td>
<td>96.8</td>
<td>92.0</td>
<td>92.4</td>
<td>71.9</td>
<td>95.8</td>
<td>84.3</td>
</tr>
<tr>
<td>Wei e et al.</td>
<td>98.0</td>
<td>82.2</td>
<td>86.1</td>
<td>85.8</td>
<td>77.0</td>
<td>95.0</td>
<td>98.3</td>
</tr>
<tr>
<td>Our Model</td>
<td>97.8</td>
<td>88.9</td>
<td>85.5</td>
<td>90.4</td>
<td>91.4</td>
<td>91.7</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Leeds Sports Pose Dataset
Deep Learning the City
Quantifying Urban Perception at a Global Scale
Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, Cesar A Hidalgo

Key Contributions
- Place Pulse 2.0 Dataset: Contains 1.17 million pairwise comparisons for 110,988 images from 56 cities, provided by 81,630 online volunteers
- Six perceptual attributes: Safe, Lively, Boring, Wealthy, Depressing, and Beautiful
- Deep network that minimizes a joint classification + ranking loss to accurately predict perception of urban attributes

The Place Pulse 2.0 Dataset

Goal: Quantify the perception of urban environments
- Helps study the relationship between a city’s physical appearance and the behavior and health of its residents
- A global dataset of human judgments in the form of pairwise comparisons of urban appearance
- Siamese-like networks, Streetscore-CNN (SS-CNN) and Ranking SS-CNN, to predict pairwise comparisons

Performance Analysis
- SS-CNN: We calculate the % of pairwise comparisons in test set predicted correctly by
  1. Softmax of output neurons in final layer
  2. Comparing TrueSkill scores [2] obtained from synthetic pairwise comparisons from the CNN
  3. Extracting features from the penultimate layer of the CNN and feeding pairwise feature representations to a RankSVM [3]
- RSS-CNN: We compare the ranking function outputs for both images in a test pair to decide which image wins, and calculate the binary prediction accuracy.

Prediction Performance Across Attributes

Table 1: Pairwise comparison prediction accuracy for standard networks fine-tuned with the Place Pulse 2.0 dataset. RSS-CNN (VGGNet) obtains the best performance

Visualizations

Figure 3: Example results from Place Pulse 2.0 dataset.

References

Figure 1: User Interface for Crowdsourced Online Game

Figure 2: SS-CNN and RSS-CNN (with additional layers in light blue)

Our dataset contains a set of $m$ images $I = \{x_i\}_{i=1}^m$ and set of $N$ comparison triplets $P = \{(i_k, j_k, y_k)\}_{k=1}^N$, $i, j \in \{1, \ldots, m\}$, $y \in \{+1, -1\}$. We want to learn a ranking function $f_r(x)$ that satisfies the maximum # of constraints [1, 3]

$$y \cdot (f_r(x_i) - f_r(x_j)) > 0 \ \forall \ (i, j, y) \in P$$

To solve this problem, we construct networks (1) SS-CNN that minimizes a classification loss $L_c$, and (2) RSS-CNN that minimizes a ranking loss $L_r$, in addition to $L_c$.

$$L_c = \sum_{(i,j,y) \in P} -y \log(g_{\text{softmax}}(y \cdot f_r(x_i), y \cdot f_r(x_j)))$$

$$L_r = \sum_{(i,j,y) \in P} \left( \max(0, y \cdot (f_r(x_i) - f_r(x_j))) \right)^2$$

RSS-CNN is trained with a joint loss $L = L_c + \lambda L_r$, where $\lambda$ is set using grid-search to maximize classification accuracy.
Motivation
Existing techniques for 4D dynamic scene reconstruction suffer from the following limitations:
1. Assume a full reconstruction of object surface at each time frame;
2. Sequential alignment of partial surfaces suffers from errors due to drift and failure for rapid and complex motion [1].

Contributions
Contributions of this work:
1. 4D Match Trees for non-sequential global alignment of partial reconstructions of non-rigid shape from single or multiple-view [3] sequences;
2. Sparse wide-timeframe matching between image pairs of non-rigid shape using a segmentation-based feature detector [2].

4D Match Tree framework for global alignment of partial surface reconstructions

Acknowledgement
The research was supported by European Commission, FP7 IMPART project (grant 316664).

References
EIGEN APPEARANCE MAPS OF DYNAMIC SHAPES
Adnane Boukhayma Vagia Tsiminaki Jean-Sébastien Franco Edmond Boyer
LJK, Université Grenoble Alpes, Inria Grenoble Rhône-Alpes, France

Context
Objective
Building efficient appearance representations for multi-view 4d models.

Issues
High dimensionality of the captured appearance information. Appearance misalignments due to geometry proxy.

Contribution
PCA based representation accounting for both appearance change and local geometric inaccuracies.

Strategy
Building representation
All views, over space and time, are used to build Eigen Maps:
• Eigen Textures for appearance variation.
• Eigen Warps for local geometric inaccuracies.

Using representation for texture generation

Quantitative evaluation
Compactness
SSIM reconstruction error w.r.t. Eigen space dimension.

Generalization ability
SSIM reconstruction error w.r.t. training set size.

Applications
Interpolation
Merging a source and a target pose geometry and appearance by interpolating in texture space and warp space separately.

Completion
Filling in missing appearance information by taking benefit of all observations accounted for in the representation.

Dataset
THOMAS
207 frames of running, walking, turning left and right.
2048x2048 pixels per view, 68 input views per frame, 50 frames per second.

CATHY
290 frames of jumping far, close, high and low

Real-time RGB-D Activity Prediction
by Soft Regression

Jian-Fang Hu1, Wei-Shi Zheng1*, Lianyang Ma2, Gang Wang2, and Jianhuang Lai1

1Sun Yat-sen University, China, 2Nanyang Technological University, Singapore

1. Introduction

The Problem & Motivation
- Predict ongoing activities with partial activity executions.
- The progress level of ongoing activities is not available.
- Subsequences containing partial activity executions can be used to regress our activity predictor learning.

Contributions
- Explore a soft regression model, which assigns a set of soft labels to the ongoing activities.
- Formulate a new RGB-D activity feature (“LAFF”), which can be computed recursively and fast.
- Develop an activity prediction system for predicting ongoing RGB-D activities in real-time (about 40 fps).

2. Methodology

Activity Representation: Local Accumulative Frame Feature (LAFF)
- Our LAFF feature is constructed from the frame-level feature (HOG and relative locations), with temporal structure considered.
- Accumulative Feature computation with temporal interval \([t_i, t_{i+1}]: x = \frac{1}{t_{i+1} - t_i} \sum_{t_i}^T F(t)
- LAFF construction: For a sequence of T frames, LAFF is the concatenation of the accumulative features extracted from a three-level temporal pyramid.

Activity predictor learning

Soft Regression Model
- \[
\min_{W, \alpha} \left\{ \sum_{i=1}^N \left( (W^T X_i - y_i) S \right)^2 + \frac{C_1}{2} \| W \|^2 + \frac{C_2}{2} \| \alpha \|^2 \right\}
\]
 where \(S\) is a diagonal matrix, whose values increase from 0.25 to 1.

3. Experimental Results

Datasets: Online RGB-D Action (ORGBD) and SYSU 3DHOI set.
Baselines: MSSVM(activity progress levels are given).
Parameters: \(C_1 = 5000, C_2 = 1, N = 40, S\) is a diagonal matrix, whose values increase from 0.25 to 1.

Results:

![Fig. 1: RGB-D activity prediction task.](image1)

![Fig. 2: A graphic illustration of our soft regression model.](image2)

![Fig. 3: Prediction performance on the ORGBD (left) and SYSU 3DHOI (right) sets.](image3)

![Fig. 4: Soft labels learned for the activities in the SYSU 3DHOI set.](image4)

![Fig. 5: More evaluations on the SYSU 3DHOI set.](image5)
A 3D Morphable Eye Region Model for Gaze Estimation

Erroll Wood1 Tadas Baltrusaitis2 Louis-Philippe Morency2 Peter Robinson1 Andreas Bulling3

1University of Cambridge, United Kingdom, {erroll.wood, peter.robinson} @cl.cam.ac.uk
2Carnegie Mellon University, United States, {tbaltrus, morency} @cs.cmu.edu
3Max Planck Institute for Informatics, Germany, bulling@mpi-inf.mpg.de

Abstract

Morphable models are a powerful tool, but have so far failed to model the eye accurately. We present a multi-part model that includes a new 3DMM of the facial eye region, and an anatomy-based eyeball. By fitting this model to an image, we can estimate 3D gaze robustly.

Model parameters

We parameterize our multi-part model with

\[ \Phi = \{ \beta, \tau, \theta, \iota, \kappa \} \]

Shape \( \beta \) and texture \( \tau \) are controlled with linear models. Pose \( \theta \) is defined by model transforms and procedural animation. We use ambient + directional illumination \( \iota \). We assume knowledge of camera parameters \( \kappa \).

Fitting our model

We fit our model with analysis-by-synthesis: given an observed image \( I_{\text{obs}} \), we produce a synthesized image \( I_{\text{syn}} \) that best matches it. We search for optimal parameters \( \Phi^* \) as follows:

\[ \Phi^* = \arg\min_{\Phi} E(\Phi) \]

Our energy includes \( E_{\text{img}} \) that measures the dense pixel-wise similarity between \( I_{\text{obs}} \) and \( I_{\text{syn}} \), and \( E_{\text{ldmks}} \) that regularizes our model against tracked facial feature landmarks [1]. \( \lambda \) controls their relative importance.

\[ E(\Phi) = E_{\text{img}}(\Phi) + \lambda \cdot E_{\text{ldmks}}(\Phi, L) \]

We minimize \( E(\Phi) \) using gradient descent with numerical central derivatives, solving for all parameters simultaneously.

\[ \text{Reduction in } E_{\text{img}} \text{ over 60 GD iterations} \]

Gaze estimation results


Reduction in \( E_{\text{img}} \) over 60 GD iterations

Gaze error over 60 GD iterations

Model expressiveness

Top: a comparison against a state-of-the-art CNN method [4] Bottom: more PCs lets our model fit better, and gives better gaze error.

Background Modeling and Foreground Detection

- Video analysis begins with background subtraction, that allows distinguishing foreground pixels.
- Still considerable challenges face foreground subtraction methods.
- Processing per-pixel basis from the background is not only time-consuming but also can dramatically affect foreground region detection.
- Region cohesion and contiguity need to be considered in the model.
- We assume that we can regard the image sequence to be made up of the sum of a low-rank background matrix and a dynamic tree-structured sparse matrix.
- We solve the decomposition using our approximated Robust Principal Component Analysis method that is extended to handle camera motion.
- Our contribution lies in dynamically estimating the support of the foreground regions via a superpixel generation step, so as to impose spatial coherence on these regions, and to obtain crisp and meaningful foreground regions.

Methodology

- Each frame $I_p = I_1, \ldots, I_n$ is concatenated as a column $A_i$ in matrix $A$.

  \[
  A = \begin{bmatrix}
  A_1 \\
  \vdots \\
  A_n
  \end{bmatrix}
  \]

  Video structure

  Matrix Structure

- We propose sparsity-inducing norms that can incorporate prior structures on the support of the sparsity pattern

  \[
  \min_{R^L \leq R^T \leq R} \| A \circ \tau - L - S \|_F + \lambda \| S \|_1 
  \]

  \[
  \psi = \sum_{i=1}^n \sum_{j=1}^n |S_{ij}| \| S_{ij} \|_1
  \]

- We then minimize the function for the parameters $L$, $S$, and $\tau$ at a time in alternating minimization until convergence. A good initialization can be found by pre-aligning all frames in the sequence to the middle frame, using a robust multiresolution method

  \[
  L^* = \arg \min_{R^L \leq R^T \leq R} \| A \circ \tau - L - S \|_F^2
  \]

  \[
  S^* = \arg \min_{S} \| A \circ \tau - L^* - S \|_F^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n |S_{ij}| \| S_{ij} \|_1
  \]

- To reduce the ghosting artifacts that persist during optimization procedures in most background subtraction algorithms, we propose an intuitive initialization strategy. This speeds up convergence, encourages the background to lean towards the best low-rank approximation of the static parts of the scene, and initializes the sparse part to take on high probability values for regions with highest statistical leverage scores

  \[
  \ell_i = \frac{1}{\beta} \sum_{r=1}^\eta |S_{r,i}|, \quad i = 1, \ldots, \eta
  \]

  \[
  S_{ij}^* = \begin{cases} 
  \ell_i, & \ell_i \geq \frac{1}{\eta} \\
  0, & \text{otherwise}
  \end{cases}
  \]

Results

Fig. 1: Tree-structured groups in sparsity induction, division, and discarding procedure. The size and location of the groups are not known and change from one frame to next.

Fig. 2: Ghosting effects that persist in RPCA-based methods. A contaminated background model in red regions affects the foreground segmentation in green regions. Our tandem model is able to eliminate these artifacts, without post-processing.

Fig. 3: 12R results: top row is the original image, second row is the ground truth, and the last row is our unrevised results without post-processing.

Table 1: CDNet [10] dataset: F-measure results for all the categories for the most competitive methods. Table accurate as of March 2016, with results from CDAero http://changeprogression.net/. The online chart keeps updating.

Table 2: SABSE [9] dataset: F-measure results for nine challenges; only the most competitive algorithms were included.

Table 3: 12R [11] dataset: F-measure results. We report our results without parameter tuning, although the dataset allows this.

More on This Topic

Contextual Priming and Feedback for Faster R-CNN
Abhinav Shrivastava and Abhinav Gupta
Carnegie Mellon University

**Goal**
Incorporate top-down information, feedback and contextual information in Faster R-CNN

**Contribution**
Using **Semantic segmentation** for contextually priming region proposal & object detection modules, and providing **iterative feedback** to the entire network

**Results**
Improvement across all three tasks: object detection, semantic segmentation and region proposals.

---

**Key Ingredients of a Region-based ConvNet Object Detector**

- **Contextual Priming and Feedback**: Incorporating top-down information Faster R-CNN
  - Main Contributions:
    - **Semantic segmentation** as a top-down signal for:
      - **Contextual Priming**
        - For region proposals & object detection
      - **Iterative Feedback**
        - Top-down feedback to the entire network
  - **From Fast R-CNN to Faster R-CNN**
    - **Faster R-CNN + Segmentation**
      - **Base Multi-task Model** (Base-MT)
        - **Stage-1** 77.3 69.5
        - **Stage-2** Init. mAP mIOU
      - **Ours [joint]**
        - **Stage-2 76.6 69.4**
        - **Stage-1 76.3 69.1**
      - **Improved across all three tasks: object detection, semantic segmentation and region proposals.**
  - **Experiments**
    - **Ablation Analysis: Contextual Priming**
      - **Faster R-CNN**
        - **Base-MT**
          - 75.6 65.8
        - **Priming to conv3/1**
          - 77.0 65.8
        - **Priming to conv3/5, each 3x3**
          - 77.8 65.9
      - **Faster R-CNN + Segmentation**
        - **Base-MT**
          - 76.5 69.3
        - **Feedback to conv3/1**
          - ImageNet
        - **Feedback to conv3(2,3,4)**
          - Stage-1 77.3 69.5
      - **More feedback helps when initializing with Stage-1 network (cf. unrolled self-feedback)**

---

**Human Visual Pathway**

- **Strong evidence of feedback connections**
  - **Top-down information**
  - **Contextual Priming**

---

**Can we bridge this gap between empirical results and theory?**

- **Incorporate top-down information, feedback and/or contextual reasoning in object detection**

---

**Contextual Priming and Feedback: Incorporating top-down information Faster R-CNN**

**Main Results** on standard dataset splits

**Detection results** on VOC07 detection test set: All methods are trained on VOC07 trainval and VOC12 trainval.

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mIOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>73.0</td>
<td>67.0</td>
</tr>
<tr>
<td>Faster R-CNN + Base-MT</td>
<td>74.7</td>
<td>67.4</td>
</tr>
<tr>
<td>Ours [joint]</td>
<td><strong>76.4</strong></td>
<td><strong>69.4</strong></td>
</tr>
</tbody>
</table>

**Detection results** on VOC07 test set: All methods are trained on VOC07 trainval and VOC12 trainval.

<table>
<thead>
<tr>
<th>Model</th>
<th>mAP</th>
<th>mIOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>68.4</td>
<td>67.0</td>
</tr>
<tr>
<td>Faster R-CNN + Base-MT</td>
<td>71.1</td>
<td>67.4</td>
</tr>
<tr>
<td>Ours [joint]</td>
<td><strong>72.6</strong></td>
<td><strong>69.2</strong></td>
</tr>
</tbody>
</table>

**Segmentation results** on VOC12 segmentation val set: All methods are trained on 07 trainval and 12 trainval.

<table>
<thead>
<tr>
<th>Model</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>66.7</td>
</tr>
<tr>
<td>Faster R-CNN + Base-MT</td>
<td>68.2</td>
</tr>
<tr>
<td>Ours [joint]</td>
<td><strong>69.9</strong></td>
</tr>
</tbody>
</table>

**Ranked 4th in 2016 COCO detection challenge with a single VGG16 model!**
Efficient Multi-view Surface Refinement with Adaptive Resolution Control  
Shiwei Li, Sing Yu Siu, Tian Fang, Long Quan

**Motivation**

Mesh refinement is nice! But it’s of very heavy computation

- Iterative, need repeated computation of refinement gradients.
- Triangle subdivision leads to high resolution mesh.

**Observation**

Different regions have unequal response of refinement

- Potential regions (e.g., edges, delicate structures) are strong
- Planar, untextured regions are weak

**Main idea**

Separate the surface into two regions:

- active: edges, has detailed geometry
- inactive: planar, low textured, no details

**Proposed method**

**Refinement framework:**

- Minimized reprojection error of surface $S$:
  $$ E_i(S) = \int_{x_i \in S} |R_i(x_i) - M_i(T_i(x_i), T_i(x))|dx_i $$
  
- Equivalent to 1) registering reprojection $R_i^{k+1}$ to reference $R_i^{k}$ and 2) aggregate the 2D gradient to 3D gradient for every vertex

- Finally the gradient $G$ for every surface point $p$ is:
  $$ G = \left[ \begin{array}{c} G_1(x_i) \cdot J_i \cdot \frac{d_i}{N_i^2d_i} N \end{array} \right] $$

- In terms of mesh $M$, solve for refinement gradient for vertex $v$ using barycentric coordinate:
  $$ G_i(p) = \sum k_i \phi_i G_{iM}(v) \longrightarrow A_{jk} G_{M} = G $$

**Adaptive resolution control:**

- Goal: improve the efficiency by only refining the important regions.

- Maximization of the utility function of accuracy and time_reduction:
  $$ u = u_{accuracy} + u_{time\_reduction} $$

- In terms of triangular mesh
  - the accuracy is quantified as the geometry improvement $g$: represented as the maximum squared distance from vertex to neighbor planes (borrowed from QEM [Garland et. Al. 1997])
    $$ g = \max \{ \pi \in planes | \{ p^i | v \} \} $$
  - the time_reduction is quantified as a factor of triangle area times #visible image pairs:
    $$ t = \frac{1}{2}(v_2 - v_0) \times (v_1 - v_0) \cdot (#visible\_image\_pairs(t)) $$

- Analyze the trade-off via curve, to find the best decision point $(x_{\alpha}, t_\alpha)$
  $$ u(x_{\alpha}, t_\alpha) = max \{ u_\alpha(x_{\alpha}, t_\alpha) \} $$
  where
  $$ u_\alpha(x_{\alpha}, t_\alpha) = u_{accuracy}(x_{\alpha}, t_\alpha) + u_{time\_reduction}(x_{\alpha}, t_\alpha) $$

- Graph-cut optimization makes the labeling smoother

**Results**

**Refinement effect:**

**Mesh resolution comparison:**

- Initial
  - #vertex: 2,438k / 318 sec
  - #vertex: 2,901k / 327 sec
  - without ARC
  - with our ARC

- Refined
  - #vertex: 663k / 56 sec
  - #vertex: 687k / 54 sec
  - without ARC
  - with our ARC

- More details at
Gaussian Process Density Counting from Weak Supervision

Matthias von Borstel¹, Melih Kandemir¹, Philip Schmidt¹, Madhavi Rao², Kumar Rajamani², Fred A. Hamprecht¹

1 Heidelberg University, HCI/IWR  2 Robert Bosch Engineering, Bangalore, India

Weakly Supervised Density Counting

• We introduce a novel learning setup: Learning to count objects within an image from only region-level count information.

• We devise a novel weakly supervised Gaussian Process (GP) that predicts object counts: i) reasonably accurately compared to earlier models all of which are trained with the expensive pixel-level annotations, ii) more accurately than all earlier models when they are also trained with region-level annotations.

The Weakly Supervised Gaussian Process Count Density Predictor

\[ p(u|Z) = N(u|0, K_{ZZ}) \]  
\[ p(f|u, X, Z) = N(f|Au, B) \]  
\[ p(g|f) = N(g|R_{uw}f, \beta^{-1}I) \]  
\[ p(c|g) = \prod_{b=1}^{B} N(c_b|g_b^T g_b, \alpha^{-1}) \]

where

\[ A = K_{zz}^T K_{zz}^{-1} \]  
\[ B = diag(K_{xx} - K_{zz}^T K_{zz}^{-1} K_{zx}) \]

Integrating \( f \) out leads to another GP that convolves the latent square root density on the Hilbert space:

\[ p(g|u) = \int p(g|f)p(f|u)df = N(g|R_{uw}Au, \beta^{-1}I + R_{uw}BR_{uw}^T) \]

Experiments

<table>
<thead>
<tr>
<th>Name</th>
<th>Data Sets # Images</th>
<th>Average Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedestrian</td>
<td>2000</td>
<td>29 ± 9</td>
</tr>
<tr>
<td>Synthetic</td>
<td>200</td>
<td>171 ± 64</td>
</tr>
<tr>
<td>Malaria</td>
<td>78</td>
<td>90 ± 84</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Pedestrians</th>
<th>Cells</th>
<th>Malaria Dense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Model</td>
<td>21.3</td>
<td>22.1</td>
<td>23.7</td>
</tr>
<tr>
<td>MIR Cluster Bags [21]</td>
<td>4.8</td>
<td>15.5</td>
<td>19.6</td>
</tr>
<tr>
<td>Bag-level Histogram Linear [2]</td>
<td>10.7</td>
<td>17.4</td>
<td>23.7</td>
</tr>
<tr>
<td>Random Forest [3]</td>
<td>3.6</td>
<td>10.0</td>
<td>21.1</td>
</tr>
<tr>
<td>Convolutional Neural Net [17]</td>
<td>-</td>
<td>7.8</td>
<td>-</td>
</tr>
<tr>
<td>GPMC (Unsmoothed)</td>
<td>15.8</td>
<td>21.2</td>
<td>26.2</td>
</tr>
<tr>
<td>GPMC (No Square)</td>
<td>20.3</td>
<td>8.6</td>
<td>29.9</td>
</tr>
<tr>
<td>GPMC (This work)</td>
<td>3.5</td>
<td>6.7</td>
<td>18.0</td>
</tr>
</tbody>
</table>

Conclusion

The proposed GP-based model learns a reasonably accurate density counting predictor from only region-level count labels.

This model is more successful at learning from weak supervision than all existing density counting models that vary from plain linear regressors to deep neural nets.
Region-based semantic segmentation with end-to-end training
Holger Caesar, Jasper Uijlings, Vittorio Ferrari

Semantic Segmentation Task

Fully Convolutional approaches

Problems:
1. Square receptive fields
2. Not multi-scale
3. Low resolution

Region-based approaches

Problems:
1. Not trained end-to-end
2. Most regions irrelevant
3. Loss ignores region size
4. How to pick positive/negative regions
5. Non-discriminative classes may determine the label due to order of softmax and max over regions

Our approach

- Differentiable region-to-pixel layer
- Differentiable free-form ROI pooling
- Correct order of softmax and max over regions
- Enable end-to-end training with multi-scale overlapping regions
- Solves problems of fully convolutional and region-based approaches

Results

SIFT Flow

PASCAL Context

Boundary Analysis

Code

Code for Fast R-CNN and our work available at https://github.com/nightrome/matconvnet-calvin
Fast 6D Object Pose Estimation from a Monocular Image using Hierarchical Pose Trees

Yoshinori Konishi, Y. Hanzawa, M. Kawade (OMRON Corp., Japan)
M. Hashimoto (Chukyo Univ., Japan)

0. Overview

- 6D pose estimation of texture-less and shiny objects from a monocular image
- Training uses only a 3D CAD of a target object
- We propose fast template-based pose estimation algorithm consisting of:
  1. PCOF: a 2D image feature which can cover wider range of 3D pose
  2. HPT: tree structured templates for efficient search of 6D object pose

1. PCOF (Perspectively Cumulated Orientation Features)

1) Generating many model images using randomized virtual camera positions
   (x: ±12 deg, roll: ±7.5 deg, distance: ±40 mm, total of 1000 images)

2) Making orientation histogram at each pixel and thresholding to extract binary-coded orientation and its weight

3) PCOF is robust to appearance changes caused by the changes in 3D object pose.

2. HPT (Hierarchical Pose Trees)

1) Clustering similar templates
   Some 2D projection images from different view angles are similar. Those similar templates are clustered based solely on their similarity scores.

2) Integration of clustered templates and hierarchization of integrated templates (making low-resolution templates)
   Integration: add orientation histograms at each pixel
   Hierarchization: add orientation histograms of 2x2 neighboring pixels

3) Building 3D pose search tree
   A pose search tree is built by iterating the above (clustering -> integration -> hierarchization). In pose estimation, initial pose candidates are detected by scanning the top level of image pyramid and they are narrowed down by tracing the tree down to the true pose.

3. Experimental results

Our proposed method (PCOF + HPT) was evaluated and compared with the state of the art on our 9 texture-less objects dataset. Recognition examples and recognition rate – FPPI curves are shown (4 from 9 objects).

Averaged recognition rate and processing time (FPPI=0.5)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate</td>
<td>0.756</td>
<td>0.600</td>
<td>0.704</td>
<td>0.858</td>
</tr>
<tr>
<td>Time (ms)*</td>
<td>1401.0</td>
<td>368.0</td>
<td>1177.3</td>
<td>156.2</td>
</tr>
</tbody>
</table>

* Core i7 3.4GHz (single core)

References

Overview

We propose a CNN that utilizes local (person bounding box) and global (full image) context through feature fusion to make human activity label predictions and achieves state-of-the-art accuracy on the HICO [1] and MPII [2] datasets.

The HICO dataset has multiple image-level labels, with multiple people in an image performing possibly different tasks. The MPII dataset has a single label per image, but multiple people all performing the same task.

Multiple Instance Learning

We aggregate predicted action scores over all people in the image before computing loss as per-person labels are not provided. Thus, at least one person is performing the action if the image-level label is active and none if the label is inactive.

\[
score(a_i) = \max_{i \in I} \{ \text{score}(a; n, b, I) \}
\]

Weighted Loss

To compensate for the unbalanced training set (average negative to positive ratio of 6000:1), we weight mistakes differently \((w_{n} = 10 \text{ and } w_{a} = 1)\).

\[
loss(I, B, y) = \sum_{i=1}^{\text{num_classes}} w_{n} \cdot y_{i} \cdot \log(y_{i}) + w_{a} \cdot (1 - y_{i}) \cdot \log(1 - y_{i})
\]

Results

We transfer domain-specific knowledge extracted from the above action datasets for answering fill-in-the-blank questions of the Visual MadLibs dataset, on the person's activity and person-object relationship question types.

Image features: Generic VGG-16 (fc7), Full-image action network (fc7, class logits), Early-Fusion action prediction network (fc7, class logits).

Answer Choice features: word2vec features averaged over words in choice.

We learn a shared embedding space for image features and answer labels using normalized Canonical Correlation Analysis (nCCA) and rank choices based on similarity.

References

A Software Platform for Manipulating the Camera Imaging Pipeline
Hakki Can Karaimer & Michael S. Brown
Department of Electrical Engineering and Computer Science
Lassonde School of Engineering, York University, Canada

{karaimer,mbrown}@ececs.yorku.ca

Introduction

There are a number of processing steps applied onboard a digital camera that collectively make up the in-camera imaging pipeline. Unfortunately, the imaging pipeline is typically embedded in a camera’s hardware making it difficult for researchers working on individual components to do so within the proper context of the full pipeline. This not only hinders research, it makes evaluating the effects from modifying an individual pipeline component on the final camera output challenging, if not impossible. We present a software platform to allow easy access to each stage of the in-camera image processing pipeline. The platform allows modification of the parameters for individual components as well as the ability to access and manipulate the intermediate images as they pass through different stages of the imaging pipeline.

In-Camera Image Processing Pipeline Stages

Stage 1: Reading the raw image (Parameters: None) The unmodified raw image is read from the DNG image file.

Stage 2: Black light subtraction and linearization (Parameters: Level values or 1D LUT) The unmodified raw image is linearized such that its values range from [0–1] in the processing pipeline.

Stage 3: Lens/Flat Field correction (Parameters: A 3x3 array, A = 1 means MATLAB function call) Many cameras provide a spatially varying correction that compensates for lens distortion and uneven light field.

Stage 4: Demosaicing (Parameters: func) The demosaic step converts the single channel raw image to three full-size R/G/B color channels by interpolating the missing values in the Bayer pattern.

Stage 5: Noise reduction (Parameters: func) This function (not provided in the Adobe SDK) has access to the intermediate image and returns back a filtered image to the pipeline.

Stage 6: White-balancing and color space conversion (Parameters: Two 3x3 matrices) This stage performs the necessary color space conversion between the camera specific RGB color space and a standard color space (ProPhoto RGB).

Stage 7: Hue/Sat map application (Parameters: 3D LUT) This procedure is intended to be part of the color space conversion to allow a non-linear transformation to be incorporated to improve the color rendition.

Stage 8: Exposure compensation (Parameters: EV value, 1D LUT) The exposure compensation is a digital exposure adjustment.

Stage 9: Color manipulation (Parameters: 3D LUT) Cameras often apply their own proprietary color manipulation that is linked to different picture styles on the camera. Like the Hue/Sat map, this is applied as a 3D LUT where RGB values are interpolated based on the table’s entries.

Stage 10: Tone-curve application (Parameters: 3D LUT) A camera-specific tone-map can be specified. This is part of the photo-finishing process on board the camera.

Stage 11: Final color space conversion (Parameters: A 3x3 Matrix) This color space conversion converts the internal camera working color space into the final output-reflected color space.

Stage 12: Gamma curve application (Parameters: 1D LUT) The final stage is a gamma curve that is applied as a 1D LUT.

Experiments and Results

We demonstrate the usefulness of our platform on a variety of examples. These examples help show how our platform can be useful in evaluating various components of the in-camera imaging pipeline.

Noise Reduction

For restoration problems, such as motion deblurring, the non-linear processing in the in-camera pipeline can change the point-spread-function (PSF) from spatially invariant, to one that is spatially varying [5]. This can be easily demonstrated with our software platform, where we can show the results of deblurring applied to the raw image vs. the sRGB image.

Deblurring

A motion blur is applied to the raw image and then run through the full pipeline to obtain the sRGB output. (B) Deblurring applied to the raw image and then its output using the camera pipeline, and the results obtained by directly deblurring the blurred sRGB image.

Noise Reduction

Noise reduction methods are typically applied and evaluated on the sRGB output. However, true noise reduction is only applied before the non-linear tone-mapping and color manipulation. Here we show the disadvantages of using image denoising outside the proper context of the full imaging pipeline.

Colorimetry

Stage 6 (color space conversion) is crucial in making sure that different camera-specific color spaces align to the same canonical color space after color conversion. Our software platform allows us to compare the results of the native camera’s colorimetric ability with two other approaches: 1) the widely used X-Rite calibration software and 2) a recent method by Bastani and Funt [1].

Acknowledgements

This study was funded in part by a Google Faculty Research Award 2015. We also like to thank Dr. Eric Chan from Adobe Research for his discussions on the Adobe DNG SDK.
A Benchmark and Simulator for UAV Tracking
Matthias Mueller, Neil Smith, and Bernard Ghanem
King Abdullah University of Science and Technology (KAUST)

Acknowledgements. Research reported was supported by competitive funding from King Abdullah University of Science and Technology.

**UAV Tracking Benchmark**

**Dataset**
- 123 aerial sequences, >110,000 frames
- Seven times larger than VIVID
- Second largest object tracking dataset

**Evaluation**
- Attributes (aerial tracking)
- Spatial robustness,
- Sensitivity to frame rate
- Long-term tracking

<table>
<thead>
<tr>
<th>Dataset</th>
<th>UAV123</th>
<th>UAV20L</th>
<th>VIVID</th>
<th>OTB100</th>
<th>OTB50</th>
<th>TC128</th>
<th>VOT14</th>
<th>VOT15</th>
<th>ALOV300</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences</td>
<td>123</td>
<td>20</td>
<td>9</td>
<td>51</td>
<td>100</td>
<td>128</td>
<td>25</td>
<td>60</td>
<td>114</td>
</tr>
<tr>
<td>Min frames</td>
<td>100</td>
<td>171</td>
<td>120</td>
<td>71</td>
<td>71</td>
<td>171</td>
<td>46</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Max frames</td>
<td>915</td>
<td>2934</td>
<td>1808</td>
<td>578</td>
<td>500</td>
<td>429</td>
<td>414</td>
<td>365</td>
<td>483</td>
</tr>
<tr>
<td>Total frames</td>
<td>112578</td>
<td>58679</td>
<td>16274</td>
<td>29381</td>
<td>50040</td>
<td>55366</td>
<td>10396</td>
<td>21871</td>
<td>151657</td>
</tr>
</tbody>
</table>

**Attribute Comparison**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>OTB100</th>
<th>UAV123</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV</td>
<td>64%</td>
<td>68%</td>
</tr>
<tr>
<td>ARC</td>
<td>48%</td>
<td>48%</td>
</tr>
<tr>
<td>FM</td>
<td>28%</td>
<td>28%</td>
</tr>
<tr>
<td>POC</td>
<td>73%</td>
<td>73%</td>
</tr>
<tr>
<td>POC</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>BC</td>
<td>21%</td>
<td>21%</td>
</tr>
<tr>
<td>IV</td>
<td>31%</td>
<td>31%</td>
</tr>
<tr>
<td>VC</td>
<td>60%</td>
<td>65%</td>
</tr>
<tr>
<td>CM</td>
<td>39%</td>
<td>42%</td>
</tr>
<tr>
<td>SOB</td>
<td>39%</td>
<td>42%</td>
</tr>
</tbody>
</table>

**Attribute Distribution**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>OTB100</th>
<th>UAV123</th>
</tr>
</thead>
<tbody>
<tr>
<td>SV</td>
<td>64%</td>
<td>68%</td>
</tr>
<tr>
<td>ARC</td>
<td>48%</td>
<td>48%</td>
</tr>
<tr>
<td>FM</td>
<td>28%</td>
<td>28%</td>
</tr>
<tr>
<td>POE</td>
<td>73%</td>
<td>73%</td>
</tr>
<tr>
<td>POC</td>
<td>30%</td>
<td>30%</td>
</tr>
<tr>
<td>BC</td>
<td>21%</td>
<td>21%</td>
</tr>
<tr>
<td>IV</td>
<td>31%</td>
<td>31%</td>
</tr>
<tr>
<td>VC</td>
<td>60%</td>
<td>65%</td>
</tr>
<tr>
<td>CM</td>
<td>39%</td>
<td>42%</td>
</tr>
<tr>
<td>SOB</td>
<td>39%</td>
<td>42%</td>
</tr>
</tbody>
</table>

Results on OTB100, UAV123 and UAV20L

**Simulator (Unreal Engine 4)**

**Highlights**
- UAV Physics Simulation
- Visual servoing system
- Frame capture and flight logging
- MATLAB/C++ integration of trackers

**Synthetic Sequence Generation**
- Custom depth maps for any mesh/object
- Automatic ground truth annotation

**Live Tracking with Feedback**
- Planned path or manual control of target
- UAV is controlled by tracking algorithm
- Live visual feedback and novel evaluation

**Qualitative Visualization**
- Generate UAV trajectories from log files
- User-defined camera views
- VR integration with HTC Vive

Benchmark and Simulator available at: https://goo.gl/LBC4zU
Scene Depth Profiling Using Helmholtz Stereopsis

Hironori Mori, Roderick Köhle, Markus Kamm
Stuttgart Technology Center, Sony Europe Limited, Germany

Introduction

- Helmholtz stereopsis is a 3D reconstruction technique to capture surface depth independent of the reflection properties of the material.
- It captures a pair of reciprocal images by exchanging the position of light source and camera.
- The resulting image pair relates the image intensities and scene depth profile by a partial differential equation.

- Boundary condition and noise propagation are fundamental key issues.
- Priors are used to solve the PDE problems by dynamic programming, producing a global matching of points in an epipolar line [1][2].

Contributions

- We developed a novel depth estimation method to utilize the integration-based representation of the Helmholtz condition.
- We propose to limit the illumination angle of the light source, such that only mutually visible parts are projected, resulting in stable boundary conditions.
- We propose a fast and accurate depth reconstruction method as it does not impose any regularizing constraints on the surface.

Integral Helmholtz condition

- Consider a surface patch mutually visible in both reciprocal images. The area integral of the normalized intensities is identical:

\[ \int_{\Omega_1} J_1(u, v) \, d\Omega_1 = \int_{\Omega_2} J_2(u, v) \, d\Omega_2 \]

- In case of rectified stereo images, the cumulative intensity along an epipolar line forms a hysteresis curve. The disparity is determined by measuring the lag between both curves.

- Noise propagation is reduced by matching the cumulative intensity at the boundaries.

Boundary condition

- The solution of the differential equation depends on the boundary conditions provided by the scene.
- By using a light projector with a matched aperture we are able to provide stable boundary conditions.
- It can be also realized e.g. by a hard threshold of image intensity, or using stereo correspondences.

Results

Experimental black background scene

Combination with stereo matching

Illumination reticles

Simulation of different material properties

Discussions

- An open problem is:
  - Occlusion and visibility order
    - The shadow cast by an occluding object may cause a change of object order between reciprocal image pairs.

- The error may be reduced
  - by the fusion with stereo matching.
  - in combination with structured light or fringe projection methods.

References

1. Paper outline

We apply the Variable Projection (VarPro, §3) method, recently shown to be effective in matrix factorization [2], to the related problem of bundle adjustment (BA) for uncalibrated cameras (§2). We present a taxonomy of VarPro-based BA algorithms by unifying the affine and the projective camera models (§3). Assuming cameras and points are initialized arbitrarily, we discover that:

- VarPro performs well on Affine BA as expected but not on Projective BA, and
- a two-stage meta-algorithm (TSMA, §4b) – VarPro on Affine BA, then VarPro on Projective BA – widens the convergence basin (§6).

2. Bundle adjustment (BA) for uncalibrated camera models (using a unified notation)

Given some 2D point tracks, find a set of camera parameters \( \{P_i\} \) and a 3D structure \( \{x_i\} \) which minimizes

\[
\sum_{i,j} \|P_i(x_{ij} - m_{ij})\|^2
\]

where \( m_{ij} \in \mathbb{R}^2 \) is the observed projection of point \( j \) in frame \( i \),

\[ P_i = P(p_i, q_i, s_{ij}, c_i) = \begin{bmatrix} P_{x1} & P_{x2} & P_{x3} & P_{x4} \\ P_{y1} & P_{y2} & P_{y3} & P_{y4} \end{bmatrix} \]

is the camera matrix at frame \( i \),

\[ x_{ij} = \tilde{x}(x_i, t_j) = [x_{i1}, x_{i2}, s_{ij}, t_j] \]

is the homogeneous representation of point \( j \),

\[ \pi(x, y, z) = [x/z, y/z, 1]^T \]

and \( \Omega \) denotes the set of visible observations.

3. The Variable Projection Method (VarPro)

**a. Linear VarPro** [1]

VarPro is a method designed for solving separable nonlinear least squares, e.g.

\[
\min \|\alpha(u,v)\|^2 = \min \|\alpha(u) - m\|^2
\]

where \( m \) is the measurement vector, \( u \) and \( v \) are the model parameters and \( \alpha(u) \) is a matrix which depends on \( u \) only.

Since \( \alpha(u, v) \) is linear in \( v \), we can obtain a closed-form minimizer for \( v \):

\[
\nu = \min \|\alpha(u, v) - m\|^2
\]

VarPro uses a Newton-like optimizer to solve the reduced problem

\[
\min \|\alpha(u) - m\|^2
\]

Depending on the approximation used for the Hessian, we have the following algorithms [3]:

- RW1, which uses the Gauss-Newton matrix,
- RW2, which uses an approximated Gauss-Newton matrix [3], and
- RW3, which is alternation [2].

**b. Nonlinear VarPro** [4]

Streloc extends VarPro to solving non-separable nonlinear least squares. e.g. \( \min \|\alpha(u, v)\|^2 \), using an iterative optimizer:

1. Initialize the residual \( r(u,v) \) using \( \nu \),
2. Linearize the residual \( r(u,v) \) to obtain a Jacobian \( J(u,v) \),
3. Solve for \( u \) using \( \alpha(u) = m \)
4. Repeat for next \( v \):
5. Repeat for next \( u \):

We propose the RW counterparts for the nonlinear extension of VarPro.

4. Proposed algorithms

None of the Projective BA algorithms succeeds from arbitrary cameras and points. VarPro-based algorithms for Affine BA have a large convergence basin.

Hence, we propose to first perform Affine BA using VarPro then warm-start Projective BA.

**a. BA algorithms for uncalibrated camera models**

<table>
<thead>
<tr>
<th>ID</th>
<th>Camera Type</th>
<th>Form</th>
<th>Strategy</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARHW</td>
<td>Affine Homogeneous Nonlinear VarPro</td>
<td>RW1</td>
<td>Projective Homogeneous Nonlinear VarPro</td>
<td>RW2</td>
</tr>
<tr>
<td>ARHW2P</td>
<td>Affine Homogeneous Linear VarPro</td>
<td>RW2</td>
<td>Projective Homogeneous Nonlinear VarPro</td>
<td>RW1</td>
</tr>
<tr>
<td>PHRW1P</td>
<td>Projective Homogeneous Nonlinear VarPro</td>
<td>RW1</td>
<td>Projective Homogeneous Joint Optimization</td>
<td>LM</td>
</tr>
</tbody>
</table>

**b. Two-stage meta-algorithms (TSMA)**

<table>
<thead>
<tr>
<th>ID</th>
<th>First-stage (affine) algorithm</th>
<th>Second-stage (projective) algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSMA1</td>
<td>ARHW2P</td>
<td>PHRW1P</td>
</tr>
<tr>
<td>TSMA2</td>
<td>ARHW2P</td>
<td>PHRW1P</td>
</tr>
<tr>
<td>TSMA3</td>
<td>ARHW2P</td>
<td>PHRW1P</td>
</tr>
<tr>
<td>TSMA4</td>
<td>ARHW2P</td>
<td>PHRW1P</td>
</tr>
</tbody>
</table>

References


5. Datasets

**a. Synthetic datasets**

<table>
<thead>
<tr>
<th>ID</th>
<th>Dataset</th>
<th>( d )</th>
<th>sce</th>
<th>Mert</th>
<th>Me</th>
<th>H</th>
<th>( t )</th>
<th>( x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0</td>
<td>30.0</td>
<td>No</td>
<td>76.92</td>
<td>2.947244</td>
<td>0.862539</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S10L</td>
<td>10.5</td>
<td>Yes</td>
<td>77.61</td>
<td>34.871149</td>
<td>0.863867</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S13L</td>
<td>13.0</td>
<td>Yes</td>
<td>77.61</td>
<td>25.100919</td>
<td>0.863871</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S11L</td>
<td>11.0</td>
<td>Yes</td>
<td>76.92</td>
<td>7.261506</td>
<td>0.862520</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S10.5L</td>
<td>10.5</td>
<td>Yes</td>
<td>77.61</td>
<td>25.100919</td>
<td>0.863871</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**b. Real datasets**

<table>
<thead>
<tr>
<th>ID</th>
<th>Dataset</th>
<th>( d )</th>
<th>sce</th>
<th>Mert</th>
<th>Me</th>
<th>H</th>
<th>( t )</th>
<th>( x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>S10.5L</td>
<td>S10.5</td>
<td>S13L</td>
<td>S12L</td>
<td>S11L</td>
<td>S11.0</td>
<td>S10.5L</td>
<td>S10.5L</td>
<td>S11.0</td>
</tr>
<tr>
<td>VGG</td>
<td>190 411</td>
<td>60.73</td>
<td>2.750877</td>
<td>0.441660</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoCo</td>
<td>196 827</td>
<td>80.71</td>
<td>0.530633</td>
<td>0.489283</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merton</td>
<td>5 1331</td>
<td>54.64</td>
<td>3.424812</td>
<td>0.135711</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. Results and conclusion

- Each run is initialized from arbitrary cameras and points.
- We evaluate 100 runs per dataset per two-stage meta-algorithm (TSMA, §4).
- TSMA1 and TSMA2 return global optimum in a large fraction of runs on most datasets.
- The convergence basin can be greatly enhanced using the right combination of methods.

Acknowledgement

The work was supported by Microsoft, Toshiba Research Europe and JNE Systech.

The conference travel was supported by the British Machine Vision Association (BMVA), Cambridge University Engineering Department (Ford of British Trust) and Christ’s College (University of Cambridge).

Project repository

https://github.com/jhh37/projective-ba/

* Much of the work was done while the first author was an intern at Toshiba Research Europe.
Localizing and Orienting Street Views Using Overhead Imagery
Nam Vo and James Hays, Georgia Tech

1. Problem formulation: image ranking
A new large scale cross view dataset: 11 cities, more than 1m pairs

2. Deep learning approach
Classification: input both images and classify whether they are a match or not
Classification Alexnet & Siamese-classification hybrid network

Representation learning: embed both images in the same feature space and compare with Euclidean distance.
Siamese and triplet (ranking) network

Distance based logistic (DBL) layer for Pair and Triplet

Contrastive loss vs log loss with DBL for pair
Hinge loss for triplet vs log loss with DBL for triplet

3. Rotational invariant matching
As orientation alignment is not available during testing, the desired representation should be invariant to rotation of overhead image, which is achieved by:
- Data augmentation by random rotation
- Better matching with multi-rotation testing and feature averaging
- Learning better representation with orientation regression

4. Experiment
Architecture comparison:

Rotational invariance:
Rotation invariance achieved with different training settings
1 rotation crop
4 rotation crops
16 rotation crops
GT (aligned overhead image)

Exhausting mini-batch sampling: improve performance & convergence rate

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-13-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.
Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.
Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, Abhinav Gupta

- How do we collect realistic and diverse videos in various homes?
- Not everything is available online
- Can we create content on AMT?

Samples of words:
- Kitchen: vacuum, groceries, chair, refrigerator, pillow
- Laughing while drinking and putting washing closing

- Recorded Videos:
  - A person is taking a picture from a fridge, then opens the fridge and begins drinking out of a jug of milk before closing it.
  - A person stands in the kitchen and cleans the fridge. Then starts to put groceries away from a bag.

Dataset for unstructured video activity recognition and commonsense reasoning for daily human activities:
- 9,848 annotated videos
- 66,500 localized activities
- 157 activities defined from verbs+nouns
- 30.1 seconds on average

Baseline results:
- Numerous baselines for benchmarking activity understanding
- Significant challenges:
  - Sequences of overlapping activities
  - Fine-grained categories
  - Actor and scene diversity
- Mean average precision over each activities in a video

Baseline Results

Examples:
- Playing TV
- Watching laptop
- Drinking coffee
- Taking phone
- Picking up bag

Collection statistics:
- $81 per video
- 1225 videos collected per day
- 267 different users participated

Sign-up bonuses:
- $1 per video
- 1000 points per video
- 2000 points per video
- 3000 points per video
- 5000 points per video
- 7000 points per video
- 10000 points per video
- 15000 points per video
- 20000 points per video
- 25000 points per video
- 30000 points per video
- 35000 points per video
- 40000 points per video
- 45000 points per video
- 50000 points per video
- 55000 points per video
- 60000 points per video
- 65000 points per video
- 70000 points per video
- 75000 points per video
- 80000 points per video
- 85000 points per video
- 90000 points per video
- 95000 points per video
- 100000 points per video

Baseline Results

CIDEr

- 0.04
- 0.05
- 0.07
- 0.14
- 0.53

Our Dataset
Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra C. Lawrence Zitnick Martial Hebert

**Overview**

**Goal**
- Use raw spatiotemporal signal from videos to learn a visual representation
- No supervised information for learning.
- Little pre-processing of input video
- Application to action recognition and pose estimation

**Key Ideas**
- Extract sequence of frames from video
- Either leave frames in their original temporal order, or shuffle them.
- Learning task: Verify if frames are in correct temporal order

**Order verification**

*Positive Tuples*

*Negative Tuples*

**Mining Tuples**

1. Compute Frame Level Optical Flow
2. Find high motion temporal window
3. Bias sampling using optical flow
4. Sample positive and negative tuples such that:
\[
\min(|f_a - f_b|, |f_e - f_d|) > \tau_{\text{min}} \\
|f_b - f_e| < \tau_{\text{max}}
\]

**Approach**

- Mine ~900k tuples from UCF101
- Initialize randomly
- Use batch-normalization
- Train for 10 epochs
- Binary Cross-entropy loss

Minimize using SGD over all the tuples

**Ablation Analysis**

<table>
<thead>
<tr>
<th>( \tau_{\text{max}} )</th>
<th>( \tau_{\text{min}} )</th>
<th>Tuple</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>15</td>
<td>60.2</td>
<td>47.2</td>
</tr>
<tr>
<td>60</td>
<td>15</td>
<td>72.1</td>
<td>50.9</td>
</tr>
<tr>
<td>60</td>
<td>60</td>
<td>64.3</td>
<td>49.1</td>
</tr>
<tr>
<td>% neg</td>
<td>% pos</td>
<td>Tuple</td>
<td>Action</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>52.1</td>
<td>38.1</td>
</tr>
<tr>
<td>75</td>
<td>25</td>
<td>72.1</td>
<td>50.9</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
<td>67.7</td>
<td>48.6</td>
</tr>
</tbody>
</table>

**Action Recognition Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Initialization</th>
<th>Mean Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>Random</td>
<td>38.6</td>
</tr>
<tr>
<td>Ours</td>
<td>50.2</td>
<td></td>
</tr>
<tr>
<td>ImageNet pre-trained</td>
<td>67.1</td>
<td></td>
</tr>
<tr>
<td>HMDB51</td>
<td>Random</td>
<td>13.3</td>
</tr>
<tr>
<td>Ours</td>
<td>18.1</td>
<td></td>
</tr>
<tr>
<td>UCF101 pre-trained</td>
<td>15.2</td>
<td></td>
</tr>
<tr>
<td>ImageNet pre-trained</td>
<td>28.5</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>29.9</td>
<td></td>
</tr>
</tbody>
</table>

**Comparison against existing unsupervised methods**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Two Close</th>
<th>Two Order</th>
<th>DrLim</th>
<th>TempCoh</th>
<th>Video Patch</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>UCF101</td>
<td>42.3</td>
<td>44.1</td>
<td>45.7</td>
<td>45.4</td>
<td>40.7</td>
<td>50.9</td>
</tr>
<tr>
<td>HMDB51</td>
<td>15.0</td>
<td>16.4</td>
<td>16.3</td>
<td>15.9</td>
<td>15.6</td>
<td>19.8</td>
</tr>
</tbody>
</table>

DrLim: Hadsell et al., 2006; TempCoh: Mobahi et al., 2009; VideoPatch: Wang and Gupta, 2015

UCF101: 101 Action classes; ~9k train, ~1k test
HMDB51: 51 Action classes; ~3k train, 1k test

**Qualitative Results**

**Nearest Neighbors from Different Videos**

<table>
<thead>
<tr>
<th>Query</th>
<th>ImageNet</th>
<th>Ours</th>
<th>Random</th>
</tr>
</thead>
</table>

**Visualizing pool5 units**

**Pose Estimation Results**

Measuring PCK and PCKh on keypoint estimation. Upper body for FLIC; Full body for MPII

<table>
<thead>
<tr>
<th>FLIC Dataset</th>
<th>Mean PCK</th>
<th>AUC PCK</th>
<th>MPII Dataset</th>
<th>Mean PCK</th>
<th>AUC PCKh@0.5</th>
<th>AUC PCKh@0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialization</td>
<td>Mean PCK</td>
<td>AUC PCKh@0.5</td>
<td>AUC PCKh@0.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>74.5</td>
<td>36.1</td>
<td>72.9</td>
<td>34.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Video Patch</td>
<td>77.1</td>
<td>42.1</td>
<td>82.8</td>
<td>43.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DrLim</td>
<td>65.2</td>
<td>27.9</td>
<td>84.3</td>
<td>41.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>84.9</td>
<td>49.6</td>
<td>87.7</td>
<td>47.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UCF pre-train</td>
<td>78.8</td>
<td>42.0</td>
<td>86.9</td>
<td>45.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet pre-train</td>
<td>85.8</td>
<td>51.3</td>
<td>85.1</td>
<td>47.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet+Ours</td>
<td>86.2</td>
<td>52.5</td>
<td>88.0</td>
<td>49.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion**

- Needs little pre-processing and no human supervision.
- Learns a powerful visual representation for human pose.
- Correct way to answer if frames are shuffled is to focus on humans.
- At par with supervised methods.
- Complementary information to supervised representations.

Code and Models: https://git.io/vPkht
Introduction:
Our task is to infer semantic edges and occlusion relationship along the edges. In other words, infer the border ownership in image.
Different from object depth order, occlusion relationship could vary between two neighboring objects, e.g.

Contributions:
1. Build a large occlusion dataset based on the PASCAL VOC 2010, called PIOD.
2. Design a new pixel-based occlusion representation and a new loss for dense prediction.
3. The first to apply deep convolutional network over occlusion estimation from a single image.

PASCAL Instance Occlusion Dataset (PIOD):
Label over PASCAL VOC 2010, 10100 images.

Class-wise diagnosis of occlusion relationship:

Occlusion representation by orientation:

In [1], binary variables are used for edge segments to specify border ownership, ‘left’ side is foreground.
In [2], pixel-based orientation representation is used, while it is quantized to 8 bins and the context for predicting is based on a local patch.
At right-side, we allow the representation to take continuous values, and can reach long range context by adopting deep features.

Orientation loss:
For edge, we follow HED\cite{3], while for orientation, we design an orientation loss.

Two stream network architecture:

Network variations:
1. HED network\cite{3]
2. Deeplab network\cite{4]

We separately train two networks and merge them with post processing by NMS.

Future work:
1. Joint with high-level instance segments, for better infer both occlusion edges and segments.
2. Better evaluation metrics by separating object vs background edges and object vs object edges.

Reference
This work is supported by NSF award CCF-1317376. and NSF STC award CCF-1231216.
RepMatch
Robust Feature Matching and Pose for Reconstructing Modern Cities
Wen-Yan Lin, Siying Liu, Nianjuan Jiang, Minh N. Do, Ping Tan, Jiangbo Lu

Pose Estimation Problem

- Urban scenes often contain large quantities of repetitive structures. This impacts SfM’s stability.
- Detecting and explicitly modeling repetitions is difficult
- **Our Goal**: A feature matcher with intrinsic robustness to repetitive structures

Match Consistency on Repetitive Structures

- Measures two basic attributes: density and spatial extent
- It is possible to separate true and false matches using match consistency [1,2]
- Repetitive structures introduce consistent but wrong matches
- However, consistency score of correct matches tends to be much higher as there are usually many alternative false positions
- By taking the most consistent matches, we can achieve a stable core set of matches which is then expanded

\[
\arg\min_w \sum_{i=1}^N C(1 - f(m_i)) + \lambda w^T G w
\]

where \( f(.) \) takes the form derived in [1,2]:

\[
f(m) = \sum_{i=1}^N w(i) \exp\left(-\frac{\|m - m_i\|^2}{\sigma}\right)
\]

Separating into true and false sets:

\[
m_j \in \{ T, \text{if } f(m_j) > \theta \} \cup \{ F, \text{otherwise} \}
\]

### Match Illustration

**All matches**

**Epipolar**

**BF**

**RepMatch**

Illustration on real images. Black dots indicate wrong matches.

**Evaluation**

Comparison with RANSAC

Comparison on Repetitive Structures

### Conclusion

Pose estimation is potentially useful in many computer vision tasks. However, it has often been considered too unstable. This paper illustrates that stable, moderate baseline pose may be practical goal. Perhaps the problem is deserving of another look. Code can be obtained at [www.kind-of-works.com](http://www.kind-of-works.com).

Convolutional Oriented Boundaries

Contributions
We provide a fast and accurate Hierarchical Image Segmentation algorithm that results in State-of-the-Art boundaries, regions, and object proposals. The core components are:

1. A Base-CNN that predicts multi-scale contours in a single pass.
2. Orientation-specialized sub-CNNs that regress the contour orientations in the same forward pass with the Base-CNN.
3. Fast algorithm that combines multi-scale contours into Ullman-like Contour Maps (UCM), based on sparse representations of the contours.

Base-CNN for contour detection

Speed-up Using Sparse Contour Representation - Timing

State-of-the-Art Boundaries and Object Proposals

Boundaries and Regions: PASCAL Context

Qualitative Results

Segmentation and Boundary Detection

Object Proposals: PASCAL and MS-COCO comparison

Contact

http://vision.ee.ethz.ch/~cvsegmentation
http://vision.ee.ethz.ch/~biw/proposals

kmavanis@vision.ee.ethz.ch
jpontutset@vision.ee.ethz.ch

twitter: @kmavanis,
twitter: @jpontutset
Superpixel Convolutional Networks using Bilateral Inceptions

Raghudeep Gadde*, Varun Jampani*, Martin Kiefel1,2, Daniel Kappler1 & Peter V. Gehler1,2

1MPI for Intelligent Systems, Tübingen; 2Bernstein Center for Computational Neuroscience, Tübingen

We propose 'Bilateral Inception' module that propagates structured information in CNNs for segmentation.

Image Conditioned Filtering Inside CNNs

This work makes two contributions for image labeling CNNs:
1. Easy to adapt image conditioned filtering within CNN architectures.
2. Recovering arbitrary image resolutions of CNN outputs.

The proposed Bilateral Inception module implements the following prior information for segmentation.
- Pixels that are spatially and photometrically similar are more likely to have the same label.

In contrast to CNN/(Dense)CRF combinations, information is propagated directly within the CNN using image adaptive filters.

Bilateral Inception Module

Bilateral Filtering:
- Given input points with features \( F_{in} \) and output points with features \( F_{out} \). Gaussian bilateral filtering an intermediate CNN representation \( z \) amounts to a matrix-vector multiplication, for each feature channel, \( c \):

\[
\tilde{z}_c = K(\theta, \Lambda; F_{in}, F_{out})z_c
\]

\[
K_{ij} = \exp(-\theta|\Lambda f_i - \Lambda f_j|)^2 \sum_j \exp(-\theta|\Lambda f_i - \Lambda f_j|)^2
\]

- Feature transformation matrix; \( \theta \): Filter scale.

The Bilateral Inception Module (BIM) is a weighted combination of bilateral filters with different scales \( \theta^1, \ldots, \theta^K \) (see Fig.2):

\[
\tilde{z}_c = \sum_{k=1}^{K} w^k \tilde{z}_c^k
\]

Bilateral filtering is modularly implemented for the reuse of intermediate computations (see Fig.3).

Input/output points need not lie on a grid.

We use superpixels for computational reasons. Also results in full-resolution output.

All the free parameters for the BI module \( w, \{\theta^k\} \) and \( \Lambda \) are learned via backpropagation.

Experiments

We insert BI modules between 1x1 convolution (FC) layers in standard CNN architectures.

\( BI(B) \) indicates BI module after FC \( B \) layer with \( B \) number of bilateral filters.

Experiments with 3 different architectures and on 3 different datasets:

Observations:
- BI modules reliably improve CNN performance with little overhead of time.
- In addition to producing sharp boundaries (like in DenseCRF), BI modules also help in better predictions due to information propagation between CNN units.
- Fast and effective in comparison to state-of-the-art dense pixel prediction techniques.

Generalization to different superpixel layouts
- BI modules are flexible in terms of number of input/output points.
- We observe that the BI networks trained with particular superpixel layout generalize to other superpixel layouts obtained with agglomerative hierarchical clustering.

Conclusions

Bilateral Inception models aim to directly include the model structure of CRF factors into the forward architecture of CNNs. They are fast, easy to implement and can be inserted into existing CNN models.

References:
Lifted formulation of dataterm and regularizer:

\[
\rho(x, u(x)) = \min \{ \nabla_j \beta \cdot \nu, \quad \nu = (E_1 \alpha - E_j \beta) \}.
\]

Convex relaxation:

\[
\int_\Omega \rho^*(x, u(x)) + \Psi^*(\nabla u(x)) \, dx.
\]
Building Dual-Domain Representations for Compression Artifacts Reduction
Jun Guo, and Hongyang Chao
Sun Yat-Sen University, Guangzhou, P. R. China

**Problem Definition**

**INPUT:** A JPEG-compressed image  
**OUTPUT:** A reconstructed artifact-free image

**Motivation: Analyze Two Domains**

Reconstruct in **DCT** Domain  
- **Pros:** no spreading of quantization errors  
- **Cons:** hard to restore high-frequency information; how to apply in CNNs is also unclear

Reconstruct in **Pixel** Domain  
- **Contrary to DCT** domain

**Approach: Deep Dual-Domain CNN**

- A **DCT-Domain Branch**  
  - Exploit DCT-domain redundancies, e.g., inter-DCT-block correlations  
  - Leverage DCT-domain priors, e.g., range of DCT coefficients

- A **Pixel-Domain Branch**  
  - Exploit spatial redundancies, e.g., patch similarity

- An **Aggregation Network**  
  - Combine two branches to generate final outputs

**Illustration of the DCT-Domain Branch**

Let $X / Y$ be the DCT coefficients of an uncompressed / compressed image, and $Q$ be the quantization table:

$$Y = \text{ROUND} \left( \frac{X}{Q} \right) \Rightarrow Y - \frac{Q}{2} \leq X \leq Y + \frac{Q}{2}$$

**Architecture of Our Approach**

Learn the residuals instead of original images, to reduce long-term memory

Both two networks are pure CNNs, consisting of convolutions and PReLUs

**Quantitative Comparisons on BSDS500**

<table>
<thead>
<tr>
<th>Quality Evaluation</th>
<th>JPEG</th>
<th>DSC</th>
<th>ARCNN</th>
<th>Ours</th>
<th>w/o DCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 PSNR</td>
<td>27.81</td>
<td>25.81</td>
<td>29.79</td>
<td>29.10</td>
<td>29.16</td>
</tr>
<tr>
<td></td>
<td>0.675</td>
<td>0.824</td>
<td>0.819</td>
<td>0.822</td>
<td>0.827</td>
</tr>
<tr>
<td>PSNR-B</td>
<td>25.10</td>
<td>26.45</td>
<td>28.73</td>
<td>28.81</td>
<td>28.81</td>
</tr>
<tr>
<td>20 PSNR</td>
<td>30.05</td>
<td>30.07</td>
<td>31.29</td>
<td>31.41</td>
<td>31.40</td>
</tr>
<tr>
<td></td>
<td>0.867</td>
<td>0.884</td>
<td>0.885</td>
<td>0.885</td>
<td>0.885</td>
</tr>
<tr>
<td>PSNR-B</td>
<td>27.22</td>
<td>30.55</td>
<td>30.83</td>
<td>30.83</td>
<td>30.84</td>
</tr>
<tr>
<td>30 PSNR</td>
<td>31.37</td>
<td>31.84</td>
<td>33.35</td>
<td>33.53</td>
<td>33.57</td>
</tr>
<tr>
<td></td>
<td>0.894</td>
<td>0.903</td>
<td>0.902</td>
<td>0.916</td>
<td>0.916</td>
</tr>
<tr>
<td>PSNR-B</td>
<td>29.53</td>
<td>31.84</td>
<td>31.94</td>
<td>31.99</td>
<td>31.99</td>
</tr>
<tr>
<td>40 PSNR</td>
<td>32.30</td>
<td>33.23</td>
<td>34.55</td>
<td>34.73</td>
<td>34.77</td>
</tr>
<tr>
<td></td>
<td>0.917</td>
<td>0.925</td>
<td>0.926</td>
<td>0.931</td>
<td>0.931</td>
</tr>
<tr>
<td>PSNR-B</td>
<td>28.40</td>
<td>32.71</td>
<td>32.78</td>
<td>32.79</td>
<td>32.79</td>
</tr>
<tr>
<td><strong>We obtained the state-of-the-art results</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Potential Extensions**

- Remove other transform-based compression artifacts, e.g., HEVC-MSP, JPEG2000, etc  
- Improve compression efficiency

**Qualitative Comparisons on Set14**

Original | JPEG | DSC | ARCNN | Ours | Ours | Ours

- Remove other transform-based compression artifacts, e.g., HEVC-MSP, JPEG2000, etc  
- Improve compression efficiency
Deep Convolutional Neural Networks (CNNs) are playing important roles in state-of-the-art visual recognition. This paper focuses on modeling the spatial co-occurrence of neuron responses, which is less studied in the previous work. For this, we consider the neurons in the hidden layer as neural words, and define a set of geometric neural phrases on top of them. The idea that grouping neural words into neural phrases is borrowed from the Bag-of-Words (BoW) model. Next, the Geometric Neural Phrase Pooling (GNPP) algorithm is proposed to efficiently encode these neural phrases. GNPP acts as a new type of hidden layer, which punishes the isolated neuron responses after convolution, and can be inserted into a CNN model with little extra computational overhead. Experimental results show that GNPP produces significant and consistent accuracy gain in image classification.

We use BigNet [49] and Wide ResNet [50] as hidden layer as neural words, and construct a set of geometric neural phrases on top of them. The idea that grouping neural words into neural phrases is borrowed from the Bag-of-Visual-Words (BoVW) model. Next, the Geometric Neural Phrase Pooling (GNPP) algorithm is inserted before the last pooling layer of each network.

• We naturally consider the data as a set of -dimensional neuron responses, which is less studied in the previous work. For this, we consider the neurons in the hidden layer as neural words, and define a set of geometric neural phrases on top of them. The idea that grouping neural words into neural phrases is borrowed from the Bag-of-Words (BoW) model. Next, the Geometric Neural Phrase Pooling (GNPP) algorithm is proposed to efficiently encode these neural phrases. GNPP acts as a new type of hidden layer, which punishes the isolated neuron responses after convolution, and can be inserted into a CNN model with little extra computational overhead. Experimental results show that GNPP produces significant and consistent accuracy gain in image classification.

Results on some small datasets
- We use AlexNet [2] as our baseline. GNPP is inserted before the last pooling layer of each network.

Results on ImageNet
- We use AlexNet [2] as our baseline. GNPP is inserted before the last pooling layer.

ACKNOWLEDGEMENTS
This paper is supported by ARO/MICrONS contract D16PC0007, ONR N00014-12-1-0883, ARK grants W911NF-15-1-0290, Faculty Early Career Development Program (CAREER), and NSF grants 1605714 and 1642920. We thank Junhua Mao, Chihang Xie and Zhuotun Zhu for discussion.
Photo Aesthetics Ranking Network with Attributes and Content Adaptation
Shu Kong1, Xiaohui Shen2, Zhe Lin2, Radomir Mech2, Charless Fowlkes1
1Department of Computer Science, University of California Irvine, Irvine, California, USA
2Adobe Research, San Jose, USA

Abstract
Real-world applications could benefit from the ability to automatically generate a fine-grained ranking of photo aesthetics.

Highlights:
1. A deep CNN to rank photo aesthetics with pairwise rank loss
2. Joint learning of meaningful photographic attributes and image content cues which help regularize the complicated photo aesthetics rating problem
3. A new aesthetics and attributes dataset (AADB) containing aesthetic scores and meaningful attributes assigned to each image by multiple human raters
4. Two sampling strategies for computing ranking loss of training image pairs for robustness in face of subjective judgment of image aesthetics
5. State-of-the-art classification performance on the existing AVA dataset benchmark by simply thresholding the estimated aesthetic scores

Fusing Attributes and Content for Aesthetics Ranking
We first train a simple model with Euclidean loss for numerical rating of photo aesthetics

$$\text{loss}_{\text{reg}} = \frac{1}{2N} \sum_{i=1}^{N} \left\| \mathbf{y}_i - \mathbf{y} \right\|_2^2$$

(a) fine-tuning with rank loss
Based on the regression net, we apply rank loss to fine-tune the network

$$\text{loss}_{\text{rank}} = \text{loss}_{\text{reg}} + \omega_{\text{a}} \text{loss}_{\text{rank}}$$

where

$$\text{loss}_{\text{rank}} = \frac{1}{N} \sum_{i,j} \max \left( 0, \alpha - \delta(\hat{y}_i \geq y_j)(\hat{y}_i - \hat{y}_j) \right)$$

$$\delta(\hat{y}_i \geq y_j) = \begin{cases} 1, & \text{if } \hat{y}_i \geq y_j \\ 0, & \text{otherwise} \end{cases}$$

(b) attribute-adaptive network
We use logistic loss to train an attribute prediction branch.

$$\text{loss} = \text{loss}_{\text{reg}} + \omega_{\text{a}} \text{loss}_{\text{rank}} + \omega_{\text{c}} \text{loss}_{\text{att}}$$

(c) attribute and content network
Similarly, we use softmax loss to train a content prediction branch whose output is used to multiplicatively gate content-specific attribute-adaptive branches. The weighted sum of scores provides the final rating.

Aesthetics & Attribute Database (AADB)
AADB images span a range of consumer and pro photos but exclude synthetic and heavily edited images.

Compared to existing datasets (e.g., AVA [23]) it is unique in having attribute labels, multiple ratings per image, and multiple images rated by each worker.

Experimental Results
We use Spearman’s rho rank correlation ($\rho$) to measure ranking performance $\rho = 1 - \frac{6 \sum d_i^2}{N(N^2-1)}$. By thresholding the rating scores, we achieve state-of-the-art classification accuracy on AVA despite never training with a classification loss.

Analyzing Model Architecture
Analysis of content-aware model on AVA dataset. Confidence-weighted gating after fine tuning out-performs branch selection and simple branch averaging.

Performance on AVA

<table>
<thead>
<tr>
<th>Method</th>
<th>$\rho$</th>
<th>AVA (%)</th>
<th>AADB (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlexNet.FE.Conf</td>
<td>0.5923</td>
<td>75.41</td>
<td>75.39</td>
</tr>
<tr>
<td>Reg</td>
<td>0.6239</td>
<td>75.33</td>
<td>75.39</td>
</tr>
<tr>
<td>Reg+Rank</td>
<td>0.6308</td>
<td>75.39</td>
<td>75.57</td>
</tr>
<tr>
<td>Reg+Rank (within-rater)</td>
<td>0.4580</td>
<td>0.4673</td>
<td></td>
</tr>
<tr>
<td>Reg+Rank + Att</td>
<td>0.391</td>
<td>0.515</td>
<td>0.4763</td>
</tr>
<tr>
<td>Reg + Att + Cont</td>
<td>0.6782</td>
<td>77.33</td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgements: This work was supported by Adobe gift fund, NSF grants DBI-1262547 and IIS-1253636.
Contributions

- Precise, correspondence-free implicit-to-implicit registration between pairs of signed distance fields (SDFs).
- Real-time frame-to-frame camera tracking on the CPU.
- Sub-minute posterior frame-to-model global pose optimization, interleaved with generating the final reconstruction.

Camera Tracking

- Frame-to-frame, runs at 18-22 FPS multithreaded on the CPU.
- 6x6 linear system in every iteration, resulting from first-order Taylor approximation around the current pose estimate.
- On average 45 times denser than point cloud registration.

Camera Tracking

\[
\begin{align*}
A &= \sum_{v_{\text{voxels}}} \nabla^2 \phi_{\text{norm}}(\xi^t) \nabla \phi_{\text{norm}}(\xi^t) \\
\mathbf{b} &= \sum_{v_{\text{voxels}}} \left( \phi_{\text{norm}}(\xi^t) - \phi_{\text{norm}}(\xi^t) + \nabla \phi_{\text{norm}}(\xi^t) \mathbf{e}^t \right) \nabla \phi_{\text{norm}}(\xi^t) \\
\frac{\partial \phi_{\text{norm}}}{\partial \xi^t} &= \mathbf{A} \mathbf{e}^t - \mathbf{b} = \mathbf{A}^{-1} \mathbf{b} \Rightarrow \xi^t = \xi^t + \mathbf{A}^{-1} \mathbf{b} \\
\end{align*}
\]

Pose Optimization

- Frame-to-model: re-register every frame to the weighted average.
- 6-element gradient descent update.
- Coarse-to-fine scheme over voxel size.

Registration Energy

Find 6 DoF camera pose \( \xi \) that best aligns two SDFs by minimizing their direct per-voxel difference:

\[
E_{\text{reg}}(\xi) = \frac{1}{2} \sum_{v_{\text{voxels}}} \left( \phi_{\text{norm}}(\xi^t) - \phi_{\text{norm}}(\xi^t) \right)^2
\]

Results: Reconstruction

Ours no ref. Ours with ref. Ours no ref. Ours with ref. Ours no ref. Ours with ref.

Results: Tracking

Quantitative evaluation of RMS, average, min, max relative and absolute translational and rotational errors on proposed dataset and TUM RGB-D benchmark [7].

Much larger convergence basin than related methods.
Handles up to 15° rotational deviation.

References

3) Kehl, W., Navab, N., Ilic, S.: Coloured Signed Distance Fields for Full 3D Object Reconstruction (ISMAR 2011)
4) Kehl, W., Navab, N., Ilic, S.: Coloured Signed Distance Fields for Full 3D Object Reconstruction (ISMAR 2014)
Knowledge Transfer for Scene-specific Motion Prediction
Lamberto Ballan\textsuperscript{1,3}, Francesco Castaldo\textsuperscript{2}, Alexandre Alahi\textsuperscript{1}, Francesco Palmieri\textsuperscript{2}, Silvio Savarese\textsuperscript{1}

\textsuperscript{1}Stanford University, \textsuperscript{2}Second Univ. of Naples, \textsuperscript{3}University of Florence

Motivation
When given a frame of a video, humans can not only interpret the scene, but also they are able to forecast the near future.
This ability is mostly driven by their rich prior knowledge about:
• dynamics of moving agents
• semantic of the scene
We exploit the interplay between these two key elements for trajectory prediction, and apply knowledge transfer to make predictions on a new scene.

Knowledge Transfer
• Retrieval-based approach that uses scene similarity to transfer the functional properties that have been learned on the training set, to a new scene
• Scene parsing: we use the scene parsing algorithm in [37] (based on SIFT + LLC, GIST, color histograms and MRF inference to refine the labeling)
• Semantic Context Descriptors: each descriptor is a weighted concatenation of the global and local semantic context components: $p = w_g + (1 - w_l) I$
  (1) global context: C-dim vector of L2 distances between the centroid of the patch and the closest point in the full image labeled as c
  (2) local context: this is a shape-context like representation which encodes the spatial configuration of nearby patches at multiple levels

Experiments
• UCLA-courtyard dataset: 6 annotated videos, 1 scene (2 views), single-class (pedestrian), scene labeled with 8 semantic classes
• Stanford-UAV dataset [28]: 21 video, 6 physical areas, 15 different scenes, multi-class (we use pedestrian and cyclist), 10 scene labels
• Evaluation Metric: Modified Hausdorff Distance (MHD)

Results: Path Prediction

(a) Input scene
(b) “Navigation map”
(c) Input and Map

Prediction Model
• The target state is defined by its position and velocity: $x_t = [P_t, V_t]^T$
• Starting from a given initial condition $x_0$, our goal is to generate a sequence of future states $x_1, x_2, \ldots, x_T$, i.e. a path $x^*$.
• The dynamic process describing the target motion is defined by:
  (1) $P_{t+1} = P_t + (O \cos \theta_t, O \sin \theta_t) + w_k$ (constant velocity model)
  (2) $V_{t+1} = \theta(\phi(P_t, V_t, M)$ (this allows non-linear behaviors)
• A Dynamic Bayesian Network exploits $M$ for path prediction

Qualitative Examples
(a) Pedestrian
(b) Cyclist

Acknowledgements
This work is funded by Toyota (1186781-31-UDARO), ONR (1165419-10-TDAUZ), MURI (1186514-1-TBCEJ), L. Ballan is supported by an EU Marie Curie Fellowship (623930).
Weakly Supervised Localization Using Deep Feature Maps

Archith John Bency\(^1\), Heesung Kwon\(^2\), Hyungtae Lee\(^2,3\), S. Karthikeyan\(^1\) and B. S. Manjunath\(^1\)

\(^1\) University of California, Santa Barbara, CA, USA \(^2\) U.S. Army Research Laboratory, Adelphi, MD, USA \(^3\) Booz Allen Hamilton Inc., McLean, VA, USA

Abstract

- Object localization aims to recognize and locate interesting objects in an image
- Ground truth image bounding boxes is difficult to obtain for large-scale datasets
- Learning to localize from Image labels (Weak Supervision) is crucial

Overview:

- Train Deep CNN classifiers from Image Labels
- Propose bounding box candidates on the final Convolutional Feature Map's spatial grid
- Better localized candidates tend to have higher classification scores
- Rank and prune candidates using beam-search

Experimental Results on Pascal VOC 2007, 2012 and MSCOCO datasets

Introduction

Weak Supervision for Object Localization

- Strong supervision for Object localization requires object-level annotations
- Annotations include bounding boxes, segmentation maps

Deep Convolutional Neural Networks

- CNNs are state-of-the-art class of techniques for image classification and object detection
- Unified feature learning and classification

Weakly Supervised Localization Using Deep Feature Maps

CNN classification and localization

- Correlation between CNN localization of object-of-interest and corresponding class scores
- Consequence of local nature of learnt convolutional filters
- Feature Maps: The output obtained by applying learnt convolutional filters and a non-linear function on data from previous layer

- The localization algorithm operates on the final conv. layer's Feature Map
  - Alexnet: \(6 \times 6 \times 256\), VGG 16: \(7 \times 7 \times 512\), in general: \(L \times L \times C\)
- Localization candidates are subsets of feature maps characterized by boxes: \(b_i = \{x_i, y_i, w_i, h_i\}\)
- For the box \(b_i\), feature map values are re-calculated as \(f_i = f(x_i, y_i, w_i, h_i)\)
- The candidates are back-projected onto image-coordinates and further localization is performed on \(L_{map}\)

Search Strategy

- Search for the best localization candidate is organized in a search tree
- The root node corresponds to the coarsest candidate, the entire image: \(b_0 = [0, 0, L, L]\)
- Children nodes are generated by reducing the width or height by one and ranked by resultant class score
- Beam-search is applied to prune low-ranking candidates
  - Number of localization candidates are kept to be tractable
  - Avoids greedy decisions

Datasets and Metrics

- Microsoft Common Objects in Context (MSCOCO): 80 object classes

- Metrics:
  - Standard IoU detection metric
  - Object localization metric \(\text{Loc}\) introduced by Goyal et al., CVPR 2015
  - Correct Localization (CorLoc)

Qualitative Results

Quantitative Results

- Localization metric results on Pascal VOC 2012 validation set:

- Localization metric results on MS COCO validation set:

- IoU Detection and CorLoc results on Pascal VOC 2007 test set:

Acknowledgements

Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-1-0053 (the ARL Network Science CTA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.
Introduction

Challenge of Person Re-identification
Large intra-class variations (i.e. illumination, pose, occlusion, camera view) lead to highly-curved manifolds in the feature space.

Motivation
- Euclidean distance is inappropriate.
- Geodesic distance is not available due to the unknown distribution.
- Like manifold learning, using local Euclidean and graphical relationship to approximate Geodesic distance for training CNN.
- Reduce the intra-class variance while preserving the intrinsic graphical structure.

Contribution
- Moderate Positive Mining: A novel positive sample selection strategy for training CNN while the data has large intra-class variations.
- A metric weight constraint in FC layer to for better generalization ability.

Our approach

Moderate Positive Mining
- Step 1. In a mini-batch with a given anchor sample, find the hardest negative sample.
- Step 2. Find the positive samples that have smaller distance than that of the hardest negative sample.
- Step 3. Mine the hardest one among these chosen positives.

Experiments

CUHK03 Validation
- Moderate positive mining improves the rank-1 accuracy by 10 percent.
- Weight constraint gains better generalization ability.

CUHK03 & 01 Performance
- See more results on VIPeR in the paper.

Metric Weight Constraint
- Implement the Mahalanobis distance in the FC layer.
- Implement the gradient.
Learning to Track at 100 FPS with Deep Regression Networks
David Held, Sebastian Thrun, Silvio Savarese
Stanford University

**Goal**
Given an initial bounding box, track a single target through a video sequence

Desired properties
- Tracks generic objects (not class specific)
- Improves performance with data
- Real-time

**Learning Image Comparison**

**Tracker Network**
- Crop from previous frame defines “target object”
- Crop from current frame defines “search region”
- Both crops are input to neural network

**Results**
- Training data:
  - 307 videos from ALOV300 [1] (removed 7 overlapping with test set)
  - 239,283 images from ImageNet Detection Challenge [2]
- Test set:
  - 25 videos from VOT 2014 Tracking Challenge [3]

Dots represent training with 14, 37, 157, and 307 videos

**Conclusion**
- Improves performance with more training videos
- Runs at 100 FPS for a single video (not batch)
- No online training or fine-tuning required
  - Learns a generic image comparison function
- Tracker regresses directly to bounding box

**Acknowledgments**
We acknowledge the support of Toyota grant 1186781-31-UDARO and ONR grant 1165419-10-TDAUZ.

**References**
1. Goal

Given a pair of handwritten documents written by different individuals, compute a document similarity score irrespective of (i) handwritten styles, (ii) word forms, word ordering and word overflow.

2. Contributions

- **IIIT-HWS**: Introducing a large scale synthetic corpus of handwritten word images for enabling deep architectures.
- **HWNet**: A deep CNN architecture for state of the art handwritten word spotting in multi-writer scenarios.
- **MODS**: Measure of document similarity score irrespective of word forms, ordering and paraphrasing of the content.
- **Applications in Educational Scenario**: Comparing handwritten assignments, searching through instructional videos.

3. Challenges

- **Segmentation of words**
- **Invariance to writer styles**
- **Invariance to word forms**

4. IIIT-HWS Dataset

- Sample word images from IIIT-HWS dataset
- Statistics:
  - Corpus Size: 9M words
  - #Fonts: 700
  - #Fonts/Class: 100
  - #Vocabulary: ~90K
- Rendered from open source handwritten fonts available over web.
- Following parameters are varied: (i) kerning level, (ii) stroke width and mean foreground and background pixel distributions.
- Gaussian filtering to smooth the final rendered image.

5. HWNet

- Formulated as word classification network.
- Initial training done on IIIT-HWS dataset.
- Fine-tuning using lower learning rates.
- Word representation: L2 normalized penultimate layers features.
- 5 Conv. and 3 FC’s.
- Input Size: 48x128
- Batch Normalization
- Soft-max loss function.
- Weight initialization: Gaussian
- Batch gradient descent with momentum.

6. Measure of Document Similarity (MODS)

- **c) Document Segmentation**
  - A simple multi-stage bottom-up approach by associating neighboring connected components using statistical features.
  - Proposed approach gives multiple word bounding box hypotheses with a high recall.

- **d) Sum of Word Matches (SWM)**

- **e) MODS Matching Algorithm**
  - Here ‘p’ and ‘q’ refers to rectangular regions from source document and candidate document respectively.
  - Matches(p, q) returns the assignments given by the Hungarian algorithm.

7. Word Spotting

- **Datasets**: IAM offline handwriting dataset and historical collection of George Washington (GW) pages.
- **Holistic word features** extracted by penultimate layer of HWNet fine-tuned on respective datasets.
- **Reduction of error rates** by 56% and 27% on IAM, GW respectively.

8. Document Image Matching

- **HW-DocSim dataset**
  - Contains 5 unique Q/A pairs and 19 participants. Total writers: 19+5
  - Four types of plagiarism [3] introduced: (i) near copy, (ii) light revision, (iii) heavy revision and (iv) non-plagiarized.
- **Evaluation Measures**
  - Verification: Area under ROC curve (AUC)
  - Graded similarity measure: Normalized discounted cumulative gain (nDCG).

- **Matching scheme evaluations**

<table>
<thead>
<tr>
<th>Method</th>
<th>Feat.</th>
<th>nDCG</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN</td>
<td>Profile</td>
<td>0.5856</td>
<td>0.5377</td>
</tr>
<tr>
<td>BOW</td>
<td>SIFT</td>
<td>0.6128</td>
<td>0.4516</td>
</tr>
<tr>
<td>SWM</td>
<td>KCSR</td>
<td>0.7968</td>
<td>0.8231</td>
</tr>
<tr>
<td>MODS</td>
<td>CNN</td>
<td>0.8844</td>
<td>0.8302</td>
</tr>
<tr>
<td>MODS</td>
<td>CNN</td>
<td>0.8569</td>
<td>0.9465</td>
</tr>
<tr>
<td>MODS</td>
<td>CNN</td>
<td>0.8993</td>
<td>0.9720</td>
</tr>
</tbody>
</table>

References

[1] Almazan et.al, PAMI’14
[2] Clough et.al, IJCAI’15
[3] Clough et.al, IJCAI’15
[4] Doetsch et.al, IJCAI’15

Project Page: http://cvit.iit.ac.in/research/projects/cvit-projects/matchdocimgs
Praveen Krishnan is supported by TCS Research PhD Fellowship. Travel grant supported by Google India.
Semantic Clustering for Robust Fine-Grained Scene Recognition

Marian George¹, Mandar Dixit², Gábor Zogg¹, and Nuno Vasconcelos²
1 Department of Computer Science, ETH Zurich, Switzerland 2 Statistical and Visual Computing Lab, UCSD, CA, United States

1 Problem

Recognize fine-grained scenes in cross-domain settings
• Fine-grained scenes share common objects
• Varying spatial configurations of objects (cluttered scenes)
• Especially true in cross-domain settings

Example: Store scenes

- bookstore
- music store
- office supplies store

2 Semantic Clustering

Exploit explicit structure in fine-grained scenes
• Semantic scene descriptor
  - Project scene images to semantic space of object occurrences
  - Convert object occurrences in scenes to scene probabilities
• Semantic Clustering
  - Cluster semantic descriptors
  - Learn a discriminative classifier for each discovered topic & combine decisions
  - Better consensus → Better generalization

3 Conditional Scene Probabilities

- Model across a range of confidence levels
  - Flexible objects arrangements in scenes across domains
  - High-level quantization
  - Imparts invariance on representation
  - Generalizes better than lower-level features
  - No spatial encoding of objects

Experimental Evaluation

Datasets

- SnapStore
  - 18 fine-grained store scenes
  - Training: web & testing: real stores
- MIT Scene 67
  - 67 indoor scenes
  - Coarse-grained & same domain
- SnapStore, SUN & Places
  - 9 store scene classes
  - Cross-dataset performance

Dataset Bias

<table>
<thead>
<tr>
<th>Training/Test</th>
<th>SUN</th>
<th>SnW</th>
<th>Web</th>
<th>Places</th>
<th>SnapStore</th>
<th>SUN</th>
<th>SnW</th>
<th>Web</th>
<th>Places</th>
<th>SnapStore</th>
</tr>
</thead>
<tbody>
<tr>
<td>SUN</td>
<td>68.7*</td>
<td>57.1</td>
<td>65.7</td>
<td>56.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SnW</td>
<td>62.7</td>
<td>71.9*</td>
<td>60.9</td>
<td>58.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web</td>
<td>64.2</td>
<td>59.2</td>
<td>67.6*</td>
<td>53.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Places</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Average classification accuracy (%)

* Same-dataset recognition accuracy (ground truth)
• performance drop > 12% when testing on phone images
• SUN and Places have very similar distributions → not suitable for domain generalization (only ~3% drop)

Comparison with State-of-the-Art

<table>
<thead>
<tr>
<th>SnapStore</th>
<th>Average classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
</tr>
</tbody>
</table>

MIT Scene 67

<table>
<thead>
<tr>
<th>Average classification accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>3.0</td>
</tr>
<tr>
<td>4.0</td>
</tr>
<tr>
<td>5.0</td>
</tr>
</tbody>
</table>

Cross-Dataset Recognition

- Semantic clustering outperforms other methods
- Clustering DeCaF performs worse than baseline DeCaF → low-level spatial maps vs. high-level semantic features
- Similarity between SUN and Places benefits DeCaF

Scene Likelihoods (OOM)
Motivation / Objective
- Humans understand each object category in relation to others, focusing on their **commonalities and difference**.

![Commonalities and Difference Diagram](image)

- **Specialization**
- **Generalization**
- **Category-specific features**

Objective
- Classification loss of the base network
- Regularization loss of generalization/specialization layer

Quantitative Result
- Classification results on CIFAR-100 dataset
- Classification results on Places-205 / ImageNet Animal dataset

Our Solution – Taxonomy-Regularized CNN
Implement generalization/specialization processes as additional **regularization layers** to leverage class structure
- Can be applied to any types of CNNs
- With small overhead:
  - < 3% increase in memory, < 15% increase in training time

Architecture
- CNNs regularized by given (or constructed) **taxonomy Τ**
- Classification loss
- Regularization loss

![Architecture Diagram](image)

(b) Hierarchical Diff-Pooling

(a) Hierarchical Min-Pooling

Qualitative Result
- Predictions for CIFAR-100 / Places-205
- Category feature maps on CIFAR-100
- Generalization/specialization feature maps on ImageNet 22K Animal
- Taxonomy of CIFAR-100 dataset discovered using our tree construction

Can we construct a CNN that generalizes / specializes?
Playing for Data: Ground Truth from Computer Games

Stephan R. Richter
TU Darmstadt

Vibhav Vineet
Intel Labs

Stefan Roth

Vladlen Koltun

1 TU Darmstadt
2 Intel Labs
3 Joint first authors

The Curse of Dataset Annotation
- Recent progress in computer vision is driven by high-capacity models trained on large datasets.
- Current semantic segmentation models appear limited by data rather than capacity.
- Creating large datasets with pixelwise labels requires large amounts of human effort.

Idea
- We explore the use of commercial video games for creating large-scale pixel-accurate ground truth data for training semantic segmentation systems.
- Modern open world games such as GTA5, Watch Dogs, and Hitman feature extensive and highly realistic worlds.
- High realism in layout of objects and environments, textures, motion of vehicles and autonomous characters, the presence of small objects that add detail, and the interaction between player and environment.

Extraction
- Challenge: Internal operation and content of off-the-shelf games are largely inaccessible.
- Our solution: We intercept the communication of a game with the underlying graphics library.

Annotation
- Group pixels that share a common combination of mesh, texture, and shader (MTS).
- Annotating MTS with their semantic class is much faster than outlining objects.
- By recognizing MTS across frames and game sessions, annotations can be propagated to all other frames of the dataset.

Dataset Analysis
- We extracted 24,966 frames with resolution of 1944x1082 pixels from GTA5.
- Labeling 98.3% of their pixel area with corresponding semantic classes took just 49 hours.

Evaluation
- Comparison of density labeled semantic segmentation datasets for outdoor scenes. We achive a three order of magnitude speed-up in annotation time, enabling us to densely label tens of thousands of high-resolution images.

References

Acknowledgements
SR was supported in part by the German Research Foundation (DFG) under the GRK 1422. Additionally, SR and SR were supported in part by the European Research Council under the European Union’s Seventh Framework Programme FP7/2007-2013/ERC grant agreement No. 306273. Work on this project was conducted in part while SR was an intern at Intel Labs.

Ground Truth from Computer Games

1 TU Darmstadt
2 Intel Labs
3 Joint first authors

The Curse of Dataset Annotation
- Recent progress in computer vision is driven by high-capacity models trained on large datasets.
- Current semantic segmentation models appear limited by data rather than capacity.
- Creating large datasets with pixelwise labels requires large amounts of human effort.

Idea
- We explore the use of commercial video games for creating large-scale pixel-accurate ground truth data for training semantic segmentation systems.
- Modern open world games such as GTA5, Watch Dogs, and Hitman feature extensive and highly realistic worlds.
- High realism in layout of objects and environments, textures, motion of vehicles and autonomous characters, the presence of small objects that add detail, and the interaction between player and environment.

Extraction
- Challenge: Internal operation and content of off-the-shelf games are largely inaccessible.
- Our solution: We intercept the communication of a game with the underlying graphics library.

Annotation
- Group pixels that share a common combination of mesh, texture, and shader (MTS).
- Annotating MTS with their semantic class is much faster than outlining objects.
- By recognizing MTS across frames and game sessions, annotations can be propagated to all other frames of the dataset.

Dataset Analysis
- We extracted 24,966 frames with resolution of 1944x1082 pixels from GTA5.
- Labeling 98.3% of their pixel area with corresponding semantic classes took just 49 hours.

Evaluation
- Comparison of density labeled semantic segmentation datasets for outdoor scenes. We achive a three order of magnitude speed-up in annotation time, enabling us to densely label tens of thousands of high-resolution images.

References

Acknowledgements
SR was supported in part by the German Research Foundation (DFG) under the GRK 1422. Additionally, SR and SR were supported in part by the European Research Council under the European Union’s Seventh Framework Programme FP7/2007-2013/ERC grant agreement No. 306273. Work on this project was conducted in part while SR was an intern at Intel Labs.
I. Problem / Motivation:

- Human Re-identification

II. Crowded Videos / Low Resolution Pedestrians

- Challenges of re-identification in crowded videos

III. Personal, Social and Environmental (PSE) Constraints

- **Personal**:
  - Appearance and preferred speed
- **Social**:
  - Grouping and Collision avoidance
- **Environmental**:
  - Transition probabilities (Destination)

IV. Optimization

- Linear and Quadratic Costs subject to:
  - Optimized using Local Search Algorithm
  - Non-convex, NP-hard
  - Initialized with Munkres on linear costs
  - Stochastic removal of possible matches
  - Followed by greedy addition

V. Experiments

- PRID Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
<th>Rank-20</th>
<th>Rank-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN100</td>
<td>8.00%</td>
<td>10.00%</td>
<td>30.00%</td>
<td>40.00%</td>
<td>57.00%</td>
</tr>
<tr>
<td>LMNN200</td>
<td>10.00%</td>
<td>20.00%</td>
<td>30.00%</td>
<td>40.00%</td>
<td>60.00%</td>
</tr>
<tr>
<td>Mahalanobis 61 + Re-ranking</td>
<td>11.00%</td>
<td>20.00%</td>
<td>30.00%</td>
<td>40.00%</td>
<td>60.00%</td>
</tr>
<tr>
<td>Non-linear ML(11)</td>
<td>17.90%</td>
<td>39.50%</td>
<td>50.00%</td>
<td>64.50%</td>
<td>-</td>
</tr>
<tr>
<td>DiscDisc7</td>
<td>19.18%</td>
<td>41.44%</td>
<td>52.18%</td>
<td>65.30%</td>
<td>84.51%</td>
</tr>
<tr>
<td>CrowdPSE (w/o topiogy)</td>
<td>21.11%</td>
<td>46.45%</td>
<td>59.98%</td>
<td>76.63%</td>
<td>98.81%</td>
</tr>
</tbody>
</table>

- Grand Central Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank-1</th>
<th>Rank-5</th>
<th>Rank-10</th>
<th>Rank-20</th>
<th>Rank-50</th>
<th>ACC(1:100)</th>
<th>F-Score (1:1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilinear</td>
<td>18.5%</td>
<td>5.48%</td>
<td>11.50%</td>
<td>23.19%</td>
<td>58.36%</td>
<td>51.00%</td>
<td>60.61%</td>
</tr>
<tr>
<td>LMNMA [8] + Re-ranking</td>
<td>6.0%</td>
<td>12.3%</td>
<td>21.9%</td>
<td>30.76%</td>
<td>71.48%</td>
<td>62.81%</td>
<td>11.16%</td>
</tr>
<tr>
<td>SDA-II [55]</td>
<td>6.0%</td>
<td>16.2%</td>
<td>23.1%</td>
<td>40.9%</td>
<td>68.56%</td>
<td>63.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>SAM [42]</td>
<td>6.0%</td>
<td>27.3%</td>
<td>42.0%</td>
<td>51.72%</td>
<td>74.48%</td>
<td>69.0%</td>
<td>26.98%</td>
</tr>
<tr>
<td>CSDk-kerne [56]</td>
<td>11.50%</td>
<td>27.38%</td>
<td>38.54%</td>
<td>50.71%</td>
<td>74.44%</td>
<td>69.49%</td>
<td>30.43%</td>
</tr>
<tr>
<td>Multifold Learning (m) [57]</td>
<td>7.3%</td>
<td>24.54%</td>
<td>37.71%</td>
<td>54.79%</td>
<td>78.99%</td>
<td>72.1%</td>
<td>30.03%</td>
</tr>
<tr>
<td>Multifold Learning (m) [57]</td>
<td>10.55%</td>
<td>34.08%</td>
<td>44.66%</td>
<td>66.53%</td>
<td>87.87%</td>
<td>86.9%</td>
<td>32.68%</td>
</tr>
<tr>
<td>CNN Features [43]</td>
<td>13.90%</td>
<td>32.45%</td>
<td>44.62%</td>
<td>62.07%</td>
<td>83.77%</td>
<td>77.79%</td>
<td>48.96%</td>
</tr>
<tr>
<td>CrowdPSE (w/ topiogy)</td>
<td>28.56%</td>
<td>81.84%</td>
<td>93.11%</td>
<td>97.87%</td>
<td>98.36%</td>
<td>99.90%</td>
<td>67.94%</td>
</tr>
</tbody>
</table>

**Acknowledgment**: This material is based upon work supported in part by the U.S. Army Research Laboratory, the U.S. Army Research Office under contract/grant number W911NF-14-1-0294.
In the training stage, we synthesize hazy images and the transmissions.

In the test stage, we generate the dehazed image using the estimated airlight and transmission map.

To train the multi-scale network, we generate a dataset with synthesized hazy images and their corresponding transmission maps based on the NYU dataset.

Effectiveness of the fine-scale network. (b) and (d) are the transmission map and dehazed result without the fine-scale network. (g) and (i) are results with the fine-scale network.

• Compute atmospheric light from the estimated transmission map

• Automatically learn haze-relevant features

Training Data Synthesis

• The training loss is used in both coarse- and fine-scale networks.

Effectiveness of Fine-Scale Network

• Training loss

\[ L(t_i(x), t_i^*(x)) = \frac{1}{q} \sum_{i=1}^{q} ||t_i(x) - t_i^*(x)||^2 \]

Performance

Goals

• Goal
  – Recover a clear image from a single hazy image

• Contributions
  – Propose a multi-scale CNN to learn effective features
  – The scene transmission map is first estimated by a coarse-scale network and then refined by a fine-scale network.
  – Analyze the differences between traditional hand-crafted features and the features learned by the CNN.

Algorithmic Overview

Atmospheric Light

• In the training stage, we synthesize hazy images and the transmissions.
• In the test stage, we generate the dehazed image using the estimated airlight and transmission map.

Feature Analysis

Quantitative Comparisons

Real Examples

References

**Introduction**

**Motivation**
- Non-uniform light intensities and exposures across observed images are a practical and common circumstance in data acquisition for the photometric stereo.
- Ex. Different light bulbs, auto-adjusted sensor exposure.

**Objective**
- Study the effects of non-uniform light intensities and sensor exposures across observed images in the photometric stereo.

### Problem definition: Photometric stereo under varying light intensity conditions

\[
\{E^*, B^*\} = \arg \min_{EB} \| M - ELB^T \|_F^2
\]

\( M \): observation matrix \((f \times p)\)
\( E \): diagonal intensity matrix \((f \times f)\)
\( L \): light vector \((f \times 3)\)
\( B \): surface normal scaled by albedo \((p \times 3)\)
\( p \): # of pixels
\( f \): # of images

**New unknown**

1. Linear joint estimation method

\[ M = ELB^T \]

\[ E^{-1}M - LB^T = 0 \]

\[ \left[ -I_p \otimes L \right] \left[ \text{diag}(m_j) \right] \left[ \text{diag}(m_p) \right] \]

Minimum condition
\[ f \geq 5 \]
\[ p \geq 3 \]

2. Factorization based method

\[ M \rightarrow (\tilde{SH}) (H^{-1} B) \]

\( \tilde{S} \): biased surface normal scaled by albedo \((p \times 3)\)

\( \tilde{B} \): biased light vector scaled by intensity \((f \times 3)\)

\( H \): linear transformation \((3 \times 3)\)

3. Alternating minimization method

\[ \{B^{(t+1)}\} = \arg \min_B \| M - E^{(t)} LB^T \|_F^2 \]

\[ \{E^{(t+1)}\} = \arg \min_E \| M - E^{(t)} LB^T \|_F^2 \]

Minimum condition
\[ f \geq 5 \]
\[ p \geq 3 \]

### Signal-to-Quantization-Noise ratio analysis

**Without saturation**

\[ \text{SNQR} = \frac{C_M}{Q_M} \]

\[ \text{SNQR} \text{ is dependent on quantization levels } Q \]

**With saturation**

\[ \text{SNQR} = \frac{C_M}{Q_M} \]

**Conclusion**

**Synthetic Experiments**

**Setting**
- 20 images under different light intensities and directions.
- Lambertian model.

**Real Experiments**

1. Auto-exposure
2. Non-uniform light source intensities
3. Mobile phone

**Contribution**
- We develop solution methods; linear joint estimation, factorization based, and alternating minimization methods.
- We further illustrate that our proposed methods can benefit from auto-exposure, with which measurements with a greater SNQR can be obtained.
- Our experiments on synthetic and real-world examples show the importance of properly handing varying light intensities and exposures.
**Visual Motif Discovery**
- **Visual motifs** are images of visual experiences significant to many people (e.g., images of informative signs, familiar social situations)
- We aim to discover visual motifs using first-person videos recorded with wearable head-mounted cameras (e.g., GoPro HERO, Google Glass)

**Technical Challenge: Developing Commonality Discovery Tailored to First-Person Videos**
- **Commonality discovery**: Finding elements (segments, objects, scenes, ...) observed across multiple images or videos (e.g., co-segmentation [Joulin+, CVPR’10], object co-localization [Tang+, CVPR’14], co-summarization [Chu+, CVPR’15])
- Commonalities are not always significant in first-person videos: Mundane moments are frequent (e.g., looking down at the ground)
- **Key question**: How to discover meaningful commonalities from first-person videos?

**Discovering Common Scenes from multiple first-person videos**
- Visual motifs are observed commonly across visual experiences of multiple people, and are distributed sparsely through each person’s visual experience
- Given a set of frames of multiple videos, discover a group of frames that satisfy inter-video similarity and intra-video sparseness

**Learning to Detect Stopping Actions for Egocentric Attention Cues**
- As a cue of significance, we focus on a commonly-occurring moment: “when people stop to acquire important visual information from a scene”
- Stopping actions can be detected by observing ego-motion of first-person videos, and can be used to constrain the clustering process

**Efficient Clustering**
- Computing $W$ becomes a bottle-neck when a large video collection is given
- If $W$ is given by $W = \text{FFT}^T$, the eigenvalues of $L' = (AF)(AF)^T$ can be obtained by solving the eigenvalue problem on $C = (AF)^T(AF)$

**Experiments**

**Navigation dataset**: 21 visual motifs recorded multiple times
- Subject walked around certain places and watched attractive objects associated with visual motifs
- 6 different environments, 44 videos in total

**Methods**
- **Ours**: Our methods with SIFT-FV/pretrained CNN scene features
- **Ours (without L')**: Our methods using L' instead of C
- **SC**: Our methods w/o egocentric attention cues
- **EgoCue**: Egocentric attention cues alone
- **TCD, COLOC**: Baseline methods [Chu+, ECCV’12; Tang+, CVPR’14]

**Visual Motif Discovery via First-Person Vision**
Ryo Yonetani (Univ. of Tokyo), Kris M. Kitani (CMU), Yoichi Sato (Univ. of Tokyo)
A Clustering Sampling Method for Image Matting Via Sparse Coding

Xiaoxue Feng, Xiaohui Liang, Zili Zhang
feng_xiaoxue@buaa.edu.cn, liang_xiaohui@buaa.edu.cn, zhangzili@buaa.edu.cn

Problem
In this paper, we present a new image matting algorithm which solves two major problems encountered by previous sampling-based algorithms. The first is that existing sampling-based approaches typically rely on certain spatial assumptions in collecting samples from known regions, and thus their performance deteriorates if the underlying assumptions are not satisfied. The second problem is that the quality of matting result is determined by the goodness of a single sample pair which causes errors when sampling-based methods fail to select the best pairs.

Materials and Methods

Samples Gathering: Clustering the foreground and background pixels respectively via a two-level hierarchical k-means clustering framework to gather a representative set of foreground and background samples.

Samples Selecting: Choosing a set of candidate samples that could better represent the true foreground and background colors of the pixel from that representative sample sets via a simple objective function.

Estimating: The proposed method capitalizes on a sparse coding to establish an objective function for generating alpha values directly from a bunch of candidate foreground and background samples.

Conclusion
A robust sampling-based image matting approach is proposed. We select samples for each unknown pixel based on both color and spatial statistic to solve the problem of missing out true samples. Moreover, based on sparse coding, we adopt a new objective function to generate alpha values from a bunch of candidate samples directly. Finally, the quality of the estimated matte is refined using a local smooth priors. Experimental results show that the proposed method achieves more robust performance than state-of-the-art approaches evaluated on a benchmark dataset.

Quantitative Evaluation

Table 1: Evaluation of matting methods on the benchmark dataset with three trimaps with respect to SAD, MSE and Gradient error metrics.

| Method | avg. overall | avg. small | avg. large | avg. user | Method | avg. overall | avg. small | avg. large | avg. user | Method | avg. overall | avg. small | avg. large | avg. user |
|--------|--------------|------------|------------|-----------|--------|--------------|------------|------------|-----------|----------|--------|--------------|------------|------------|-----------|
|        | rank | rank | rank | rank |       | rank | rank | rank | rank |        | rank | rank | rank | rank | rank | rank |
| 1. Proposed | 10.2 | 14 | 6.3 | 10.4 | 1. LNSP | 8.8 | 6.3 | 8.3 | 12.1 | 1. KL-D | 10.5 | 9.1 | 8.4 | 14.1 |
| 2. LNSP | 10.5 | 9 | 10 | 14.5 | 2. KL-D | 11.9 | 11.5 | 10.3 | 13.9 | 2. LNSP | 10.7 | 8.2 | 10 | 13.8 |
| 3. KL-D | 11.8 | 11.3 | 10.6 | 13.6 | 3. CCM | 12.2 | 12.3 | 12.1 | 9.1 | 3. CCM | 11.8 | 12.1 | 11 | 12.4 |
| 4. Compre | 13.5 | 11.3 | 13.5 | 15.8 | 4. Compre | 13.3 | 12.3 | 13 | 14.5 | 4. CCM | 13.8 | 16.5 | 13.5 | 11.5 |
| 5. IT | 13.9 | 15.4 | 13.1 | 13.1 | 5. SRV | 13.5 | 17.5 | 11.9 | 11.1 | 5. SRV | 14 | 16.6 | 15.1 | 10.3 |
| 6. CWCT | 14.3 | 14.8 | 14.8 | 13.3 | 6. CWCT | 14.6 | 14.5 | 15.4 | 13.9 | 6. CCM | 14 | 16.2 | 12.8 | 13.4 |
| 7. SRV | 14.4 | 17.1 | 14 | 12 | 7. SRV | 15.9 | 19.4 | 9.5 | 18.4 | 7. Global | 15 | 14.9 | 16.8 | 13.4 |
| 8. SparseCoded | 14.8 | 18.1 | 15.3 | 11.1 | 8. WCT | 17.1 | 15.9 | 18.4 | 17.1 | 8. Improved? | 15.3 | 16.9 | 16.8 | 12.3 |
| 10. CCM | 16.7 | 19.5 | 16.4 | 14.3 | 10. Global | 17.8 | 13.4 | 21.6 | 18.5 | 10. Proposed | 18.2 | 19 | 7.8 | 22.4 |

Visual Comparison

Figure 1 shows the visual comparison of our approach with the recent matting methods on the doll, plant and pineapple images from the benchmark dataset.

Previous sampling-based image matting methods typically rely on spatial closeness while collecting samples which would fail to generate accurate alpha matte when the true samples are not spatially close to the unknown pixels, this problem is known as missing out true samples. Figure 2 evaluates the effectiveness of the proposed sampling method in dealing with missing out true samples problem.

Acknowledgements
This work is supported by the funds of National Natural Science Foundation of China (No.61572038) and National High Technology Research and Development Program of China (No.2015AA016402). The authors would like to thank the anonymous reviewers for their insightful comments and suggestions on this work.

References
**Contributions**

**Goal:** Obtain the most annotation cost-effective supervision for semantic image segmentation.  
- Novel, cost-efficient **supervision regime** for semantic segmentation based on humans pointing to objects.  
- Extensive human study to collect point annotations for PASCAL VOC 2012, and released annotation interfaces.  
- A generic **objectness prior** incorporated directly in the loss to guide the training of a CNN.

---

**Novel supervision regime**

**Problem:** Assign one class label to every pixel in an image.

- **Training:** Standard regime = costly per-pixel annotations  
- **Levels of supervision**
  - full supervision  
  - image-level labels  
  - points  
  - squiggles  

- **Key insight:** Annotating one pixel per training image significantly improves segmentation annotation and only marginally increases the annotation cost as compared to image-level labels.

- **Loss function for point-level supervision:** We have a small set of supervised pixels, and other pixels just belong to some class in L. Incorporation into loss function: Helps correctly infer the spatial extent of objects for models trained with very few supervised pixels.

\[
\mathcal{L}_{obj}(S, P) = -\frac{1}{|S|} \sum_{i \in S} \log(P_{i|S}) - \frac{1}{|P|} \sum_{i \in P} \log(1 - P_{i|P}) + \sum_{j \in \mathcal{O}} \log(S_j) \]

**Objectness prior in CNN loss**

**Purpose of the objectness prior:** Helps correctly infer the spatial extent of objects for models trained with very few supervised pixels.

---

**Crowdsourcing point annotations**

- **AMT annotation UI**
- **Example points collected**

**Measuring the annotation times:**
- Points and squiggles: measured directly during data collection.  
- Other types of supervision: we rely on times from literature.

**Reported annotation times:**
- Image-level labels: 20.0 sec/image  
- Points: 22.1 sec/image  
- Squiggles: 34.9 sec/image  
- Full supervision: 239.7 sec/image

---

**Results on PASCAL VOC 2012 dataset [Everingham 2010]**

**Effects of point supervision + objectness:** The combined effect results in a +13% mIOU over image-level labels.

**Point supervision variations:** Multiple object instances and multiple annotators achieve only modest improvements over single points.

**Segmentation on an annotation budget:** Point supervision provides the best trade-off between annotation time and segmentation accuracy.

---

**Bibliography**


---

**What’s the Point: Semantic Segmentation with Point Supervision**

Amy Bearman\(^1\) Olga Russakovsky\(^2\) Vittorio Ferrari\(^3\) Li Fei-Fei\(^1\)

\(^1\)Stanford University  \(^2\)Carnegie Mellon University  \(^3\)University of Edinburgh

---

[http://vision.stanford.edu/whats_the_point](http://vision.stanford.edu/whats_the_point)
Fashion Landmark Detection in the Wild

Ziwei Liu*  Sijie Yan*  Ping Luo  Xiaogang Wang  Xiaoou Tang

Motivation

Problem:
• How to achieve accurate fashion image understanding?

Challenges:
• None-rigid deformations
• Larger spatial variances
• Larger appearance variances

Dataset

Fashion Alignment

Visual Results

DeepFashion dataset is available at:
Person Re-identification by Unsupervised $L_1$ Graph Learning
Elyor Kodirov, Tao Xiang, Zhenyong Fu and Shaogang Gong
School of EECS, Queen Mary University of London

- **Existing Supervised Re-ID Approaches**
  - Most are based on supervised learning [1,2]
  - They require large number of labelled images across camera views
    - Label intensive and unsalable
    - May not even be possible under an open-world setting

- **Our Unsupervised Re-ID Approach**
  - Learning cross-view discriminative features from unlabeled data
  - No human annotation required, thus much more scalable

- **Contributions**
  - We formulate a novel graph regularised dictionary learning model for Re-ID with a new robust $L_1$-norm graph regularisation term and joint graph and dictionary learning.
  - We develop an efficient iterative optimisation algorithm for non-smooth and non-convex objective function of our model.

- **Results**

---

**Unsupervised Re-ID results measured at Rank-1**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>VIPeR</th>
<th>PRID</th>
<th>CUHK01</th>
<th>CUHK03</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>15.6</td>
<td>13.9</td>
<td>10.9</td>
<td>12.5</td>
</tr>
<tr>
<td>SDALF</td>
<td>19.9</td>
<td>16.3</td>
<td>9.9</td>
<td>4.9</td>
</tr>
<tr>
<td>DLLAP</td>
<td>29.6</td>
<td>21.1</td>
<td>28.4</td>
<td>22.3</td>
</tr>
<tr>
<td>eSDC</td>
<td>26.7</td>
<td>-</td>
<td>26.6</td>
<td>7.7</td>
</tr>
<tr>
<td>CPS</td>
<td>22.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GTS</td>
<td>25.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BGG</td>
<td>21.7</td>
<td>-</td>
<td>-</td>
<td>18.9</td>
</tr>
<tr>
<td>Ours</td>
<td>33.5</td>
<td>25.0</td>
<td>41.0</td>
<td>30.4</td>
</tr>
</tbody>
</table>

**Supervised Re-ID results measured at Rank-1**

<table>
<thead>
<tr>
<th>Datasets</th>
<th>VIPeR</th>
<th>PRID</th>
<th>CUHK01</th>
<th>CUHK03</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method R@1</td>
<td>Method R@1</td>
<td>Method R@1</td>
<td>Method R@1</td>
<td></td>
</tr>
<tr>
<td>XQDA</td>
<td>40.0</td>
<td>14.5</td>
<td>34.3</td>
<td>46.4</td>
</tr>
<tr>
<td>DEEP</td>
<td>34.8</td>
<td>19.7</td>
<td>47.5</td>
<td>44.9</td>
</tr>
<tr>
<td>MLAPG</td>
<td>40.7</td>
<td>25.2</td>
<td>Deepreid</td>
<td>51.2</td>
</tr>
<tr>
<td>EKPF</td>
<td>36.1</td>
<td>16.0</td>
<td>Deepreid</td>
<td>19.9</td>
</tr>
<tr>
<td>CDEEP</td>
<td>40.9</td>
<td>18.0</td>
<td>XQDA</td>
<td>63.5</td>
</tr>
<tr>
<td>QLF</td>
<td>30.2</td>
<td>12.3</td>
<td>MLAPG</td>
<td>64.2</td>
</tr>
<tr>
<td>Ours_un</td>
<td>33.5</td>
<td>25.0</td>
<td>41.0</td>
<td>30.4</td>
</tr>
<tr>
<td>Ours_sup</td>
<td>41.5</td>
<td>30.1</td>
<td>50.1</td>
<td>39.0</td>
</tr>
</tbody>
</table>

**Acknowledgments**
This work was funded in part by the EU FP7 Project SUNNY (grant agreement no. 313243).

---

**Method**

1. **Robust $L_1$-norm graph regularisation term:**
   \[ \Omega(Y) = \sum_{i,j}^N |y_i - y_j|_1^2 w_{ij} \rightarrow ||Y A_W||_1 \]
2. **Full model:** Learn dictionary ($D$) & graph ($W$):
   \[ \min_{W,Y} \frac{1}{2} ||X - DY||_F^2 + \lambda_1 ||Y A_W||_1 + \lambda_2 ||W||_F^2 \]
   s.t. \[ ||d||_1^2 \leq 1, W_{ii} = 1, W_{ij} \geq 0, \]
   - Optimisation – we develop a solver based on the ADMM algorithm [3].

**Cross-view matching:**

- After learning $D$, given a pair of samples $x^a_i$ and $x^b_j$, we first compute their collaborative representations $y^a_i$ and $y^b_j$.
  \[ y^a_i = \arg \min_{y} ||x^a_i - Dy^a_i||_2^2 + \lambda ||y^a_i||_2 \]
  \[ y^b_j = \arg \min_{y} ||x^b_j - Dy^b_j||_2^2 + \lambda ||y^b_j||_2 \]

- Cosine distance is then used to measure visual similarity.

**Extension to supervised re-id:**
- If cross-view pair $(i,j)$ is labelled, $W_{ij}$ is set to 1 or 0.

**References**
Leveraging Visual Question Answering for Image-Caption Ranking

Xiao Lin  Devi Parikh
Virginia Tech

Overview

VQA as “feature” extraction
• VQA features $u_i$ for image $I$, $u_c$ for caption $C$

Log probabilities of a set of N (=3,000) question-answer pairs $(Q_i, A_i)$.

$$u^{(i)}_I = -\log P_{VQA,I}(A_i|Q_i, I)$$

$$u^{(i)}_C = -\log P_{VQA,C}(A_i|Q_i, C)$$

Q: What are the men wearing on their heads? A: Helmets

Q: Is it clean? A: Yes

Image-caption ranking
• Given an image, retrieve its caption from K (=1,000) captions
• Given a caption, retrieve its image from K (=1,000) images

VQA “agnostic” baseline
• Learn a scoring function $S(I, C)$ to predict whether image $I$ and caption $C$ are compatible

ImageCaption

VQA “aware” approach
• Score-level fusion
Combining VQA and baseline features at score level

ImageCaption

Result

MSCOCO-1k test

<table>
<thead>
<tr>
<th>Approach</th>
<th>Caption Retrieval Accuracy</th>
<th>Image Retrieval Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R@1</td>
<td>R@5</td>
</tr>
<tr>
<td>Baseline (Kiros et al. 2014)</td>
<td>43.4</td>
<td>75.7</td>
</tr>
<tr>
<td>VQA-only</td>
<td>37.0</td>
<td>67.9</td>
</tr>
<tr>
<td>Score-level fusion</td>
<td>46.9</td>
<td>76.8</td>
</tr>
<tr>
<td>Representation-level fusion</td>
<td>50.5</td>
<td>80.1</td>
</tr>
<tr>
<td>+ Baseline VQA (58.0% acc)</td>
<td>53.6</td>
<td>86.3</td>
</tr>
<tr>
<td>+ Better VQA (60.5% acc)</td>
<td>+10.1</td>
<td>+8.8</td>
</tr>
</tbody>
</table>

State-of-the-art on MSCOCO

VQA consistently improves image-caption ranking across different amount of image-caption ranking data

Approach

Which facts does my model want to verify for better image-caption ranking?
• The $(Q,A)$ pairs that have large mutual information between validity $V_{Q,A}$ and relevance of captions $C_i$ given image $I$

$$arg \max_{(Q,A)} I(V_{Q,A}; C_i) = arg \max_{(Q,A)} MI(V_{Q,A}; C_i)$$

Top facts that machine would like to verify

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What kind of food is this?</td>
<td>Cake</td>
</tr>
<tr>
<td>What is the brown object in the foreground of the picture?</td>
<td>Train</td>
</tr>
<tr>
<td>Where was this picture taken?</td>
<td>Beach</td>
</tr>
<tr>
<td>What activity is the man doing?</td>
<td>Skateboarding</td>
</tr>
<tr>
<td>What is the table made of?</td>
<td>Wood</td>
</tr>
<tr>
<td>What color is the train?</td>
<td>Red</td>
</tr>
<tr>
<td>Where is this place?</td>
<td>Beach</td>
</tr>
<tr>
<td>What is this person throwing?</td>
<td>Frisbee</td>
</tr>
</tbody>
</table>
DAVE: A Unified Framework for Fast Vehicle Detection and Annotation
Yi Zhou, Li Liu, Ling Shao and Matt Mellor
Northumbria University, Newcastle, UK. & Createc Ltd., Cockermouth, UK.

Motivation

Main Idea

DAVE consists of two CNNs: FVPN and ALN.
FVPN: Fast Vehicle Proposal Network (shallow fully convolutional)
ALN: Attributes Learning Network (Multi-task GoogLeNet Extension)
Joint optimization by bridging two CNNs with latent data-driven knowledge.

A Two-Stage Inference Scheme:
FVPN is performed on an input frame to predict bounding-boxes of proposals.
ALN is to implement vehicle verification, pose estimation, color recognition and type classification simultaneously.

Detection

A new UTS dataset
* six video sequences
* 1920*1080
* attributes labelled
* various illuminations and viewpoints

Other two datasets:
PASCAL VOC 2007 Car
LISA 2010 Car

Results of Attributes Annotation

Explained problems:
Single-task learning or multi-task learning?
How small a vehicle size can DAVE annotate?
Deep or shallow?
Why not an elegant One-Net pipeline?

Annotation

Results Compared to One-Net Pipeline

Training Dataset

Qualitative Results
Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input

Srinath Sridhar\textsuperscript{1} Franziska Mueller\textsuperscript{1} Michael Zollhöfer\textsuperscript{1}
Dan Casas\textsuperscript{1} Antti Oulasvirta\textsuperscript{2} Christian Theobalt\textsuperscript{1}

\textsuperscript{1}Max Planck Institute for Human Cognitive and Brain Sciences
\textsuperscript{2}Aalto University

\textbf{Tangible Interaction for AR/VR}  
\textbf{Joint Hand-Object Tracking}  
\textbf{Contributions}

\textbf{ADVANTAGES}  
- Real-time (CPU-GPU)  
- Skilled movements

\textbf{CHALLENGES}  
- Extreme occlusions of objects  
- Complex and fast motions  
- Segmentation of hands from object  
- High dimensional problem  
- Run time constraint

\textbf{NOVEL FRAMEWORK FOR REAL-TIME HAND-OBJECT TRACKING}  
- Two-camera setup  
- Real-time (CPU-GPU)

\textbf{PRACTICAL}  
- Supports various objects, shapes, and sizes

\textbf{RESULTS}  
- RESULTS ON DATA FROM TDIOWAS AL (ECCV 2016)  
- RESULTS ON DATA FROM TDIOWAS AND GALL (ECCV 2015)

\textbf{REAL-TIME TRACKING}  
- Tablet with video goes here

\textbf{Multiple Proposal Pose Optimization}  
\textbf{OBJECTIVE 1:}  
\[ E_{opt}(A) = \sum_{i=1}^{N} w_i E_i(A) = \sum_{i=1}^{N} w_i E_i(A) \]

\textbf{OBJECTIVE 2:}  
- Spatial alignment  
- Semantic alignment

\textbf{Spatial Alignment}  
- Aligns Gaussian mixtures

\textbf{Semantic Alignment}  
- Aligns GMMs with semantic label information

\textbf{CONTACT HANDLING}  
- Enforces contact between object and object Gaussians that are close

\textbf{OCCULTION HANDLING}  
- Forcibly occluded parts of the hand to move consistently with other parts

\textbf{Dexter+Object dataset available!}  
handtracker.mpi-inf.mpg.de
DeepWarp: Photorealistic Image Resynthesis for Gaze Manipulation
Yaroslav Ganin1,2, Daniil Kononenko1, Diana Sungatullina1, Victor Lempitsky1
1 Skolkovo Institute of Science and Technology, 2 Montreal Institute for Learning Algorithms

Task
In this work, we consider the task of image resynthesis. We propose a solution based on neural networks of a specific architecture that can achieve highly photo-realistic results. To demonstrate our approach, we focus on the task of gaze manipulation. Applications:
- video-conferencing
- “talking head”-type videos
- photo- and video-editing
Example: During video calls people usually look at the application window and not at the camera ⇒ gaze is directed slightly down.

Dataset
Skoltech Dataset
collection procedure:
- Special stand to minimize head movement
- Person tracks a moving point on the screen
- Various head poses and lighting conditions; eyewear
- Facial landmarks and eye region cropping

Method overview
Inputs:
- Rectangular eye region image (RGB)
- Landmarks (anchors) extracted by an external algorithm
- Correction angle α
- 25c (no cards accepted, except GPUs)

Pipeline:
- Predict coarse warping flow at half-scale
- Use half-scale output image and original input to predict final warping flow at full scale
- Apply Lightness Correction Module (LCM)

Possible training objectives:
- Regular $l_2$ (reported in the paper) or $l_1$
- Compare (using $l_2$ or $l_1$) output against multiple locations in the ground-truth and return minimum (addresses poor registration in the dataset)

Lightness Correction Module
Warping struggles with
- Subtle photometric changes
- Dis-occlusions
To address this, we propose a post-processing operation called lightness correction. Given a pre-warped image $O$, we compute the final output at pixel $(x,y)$ for channel $c$ as
$$O(x,y,c) = (1 - M(x,y)) + M(x,y) \cdot O(x,y,c)$$
where $M$ is produced by one more NN.

Quantitative results
Ordered errors for 15° redirection

Errors for different correction angles

Dataset splits:
(# persons)
- Development: 26
- Test: 7

Models:
- RF: Random forest
- SS: Single-scale
- MS: Two scales
- CFW: Two scales w/ coarse-to-fine
- +LCM: CFW w/ LCM processing

Qualitative results
15° vertical redirection

Summary
- A method for realistic gaze redirection
- New fully differentiable pipeline based on warping
- Supports arbitrary angles and directions out of the box
- Approach is general and can be applied to other tasks (WIP: DeepWarp + GAN + weak/no labels)

Reasonably fast on the GPU (can be made even faster by stripping down the architecture)

Project page (more awesome stuff!): http://bit.ly/2bJEbCy/
Given: A set of non-rigid shapes with their corresponding class labels.

Goal: Efficient estimation of similarity between labeled and new unlabeled shapes.

- Find a shape descriptor space where classification is easy.
- Efficiently compute a global shape descriptor for any new shape.
- Accurately determine the class label of the new shape.

Contributions

- An intuitive way to measure similarity between non-rigid shapes
- Small set of shapes needed for training, compared to other methods
- State of the art classification accuracy and runtime

Shape Analysis Tools

Laplace-Beltrami Operator

- Invariant under isometric deformations
- Defined as the negative divergence of the gradient: \( \Delta f = - \text{div} \nabla f \)
- LBO Eigen-decomposition of a shape, \( \Delta f = \lambda \phi \), with \( 0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_\infty \)

Heat Kernel Signature (HKS) [1]

- The amount of heat that remains at a point \( i \) after diffusion time \( t \) is \( k_t(i, x) = \sum_{k=0}^{\infty} e^{-\lambda_k t} \phi_k(x) \)
- Concatenating for different time stamps: \( \text{HKS}(x) = [k_t(x, x), k_t(x, y), \ldots, k_t(x, z)]^T \)

Scale Invariant Heat Kernel Signature (SI-HKS) [2]

- Observation: The HKS of a scaled shape is only shifted in time.
- Taking the Fourier transform of the HKS, undoes the phase and sampling at \( t \) frequencies:
  \( s\text{HKS}(x) = \langle |\text{HKS}| \rangle_{\text{complex}} \)

Wave Kernel Signature (WKS) [3]

- The avg. probability over time to locate a quantum particle at point \( x \) is:
  \( p(x) = \lim_{t \to \infty} \frac{1}{t} \int_0^t \langle \psi(x) | \psi(x) \rangle dx = \sum_{k=0}^{\infty} \phi_k(x) \phi_k(x) \)
- Different choices of \( f \) give us shape properties at different scales:
  \( \text{WKS}(x) = [p_0(x), \ldots, p_n(x)]^T \)

Large Margin Nearest Neighbor [4]

- Finds a Mahalanobis distance that is optimal for k-nearest neighbors classification.
- Enforces maximum margin between intra-class and inter-class samples (as in SVMs).
- The learned descriptor space can be expressed as a linear transformation \( \lambda \) of the inputs.
- A convex formulation of the objective function guarantees the optimal metric \( M = \lambda \).

References

Multi-Task Zero-Shot Action Recognition with Prioritised Data Augmentation

Xun Xu1,2
ellexuxu@nus.edu.sg

Queen Mary University of London, UK

Timothy Hospedales1,3
t.hospedales@qmul.ac.uk

National University of Singapore, Singapore

Shaogang Gong1
s.gong@qmul.ac.uk

University of Edinburgh, UK

1. Motivation

Motivation
Dimensions of semantic embedding are correlated
MTE better than RR
Latent Matching Embed
Discover a lower-dimensional latent semantic embedding space for more effective matching.

Category KLIEP

9

43.5

Aug

9

BoW

i

Y

9

44.3

X X

41.1

Olympic Sports

Visual KLIEP

23.4

Haircut

Uniform

strategies.
MTE model outperforms

Infer based on existing seen action videos, e.g. “Hammer Throw”, “Discus Throw”.

3. Importance Weighting for Data Augmentation

• 1. Construct a semantic embedding space (SES) with word-vectors [2].
• 2. Learn a visual-to-semantic mapping by regression.
• 3. Project videos to SES & NN match unseen class

Existing Approaches

Can we recognize an unseen action, e.g. “Hammer Throw”, “Discus Throw”.

3. Project videos to SES & NN match unseen class (prototypes).

Limitations of Existing Methods

• Dimensions of semantic embedding are correlated • single task regressors overfit.
• All training data are equally considered, risking negative knowledge transfer.

Our Model

• Multi-Task Embedding: Develop a multi-task regression
• Data Augmentation: Augment with additional datasets, e.g. HMDB-HMDB+UCF+Olympic.

Importance Weighting: Reweight augmented auxiliary data to focus on relevant examples

Contributions

• Formulate the visual-to-semantic mapping as a multi-task regression problem.
• Discover a lower-dimensional latent semantic embedding space for more effective matching.
• Selectively re-weight auxiliary data to prioritize most relevant training data.

2. Multi-Task Embedding (MTE)

Single-Task Regression
Learn visual-to-semantic mapping as regularized linear regression.

W = ZX + (XX + λI)−1 (XZ + λA)

Multi-Task Regression
Formulate the visual-to-semantic regression W as a combination of a collection of regressors A.

W = ZS + (XX + λA)−1 (XZ + λA)

Latent Matching
Conduct nearest neighbor matching in latent (L) space rather than semantic embedding (Z) space.

z∗ = arg minz′ |S(S)−1S′Z′ − AX|

Merits of Multi-Task Embedding

Unsupervised; A lower dimension latent representation L for better NN classification

3. Importance Weighting for Data Augmentation

General Idea
• Augment auxiliary set with additional dataset (Data Augmentation).
• Higher weight is given to auxiliary data-class which are more related to target testing data-class (Importance Weighing).

Kullback-Leibler Importance Estimation Procedure (KLIEP)

Objective: Estimate a weighting function σ(X) to minimize the discrepancy between augmented auxiliary distribution pθ(X) and target data distribution pθ(x).

min DKL (pθ(x) || pθ(x)) = E pθ(x) log pθ(x) σ(X) pθ(x) dx

This objective is solved by interior point methods.

Integrate the Weight with Auxiliary Data

Estimated weights can be integrated with auxiliary data by simply multiply the root square:

z = z′/σ, s = s′/σ

4. Experiment

Datasets
Three human action datasets – HMDB51 (51 classes; 6766 videos), UCF101 (101 classes; 13320 videos) and Olympic Sports (16 classes; 783 videos).

Data Split
Half-half split by half classes for train and half for test (no overlap). Five random half-half splits for evaluation.

Multi-Task (MTL) vs. Single-Task Embedding

ZSL Model

Latent Embedding

Multi-Task (MTL) vs. Single-Task Embedding

RR [3]

–

X

18.4±2.1

14.5±1.9

40.9±10.1

RMTL [4]

✓

X

18.5±2.1

14.6±1.1

41.1±10.0

RMTL [4]

✓

X

18.7±1.7

14.7±1.0

41.1±10.0

GOMLT [5]

✓

X

18.5±2.2

13.1±1.5

43.6±8.8

GOMLT [5]

✓

X

18.9±1.0

14.9±1.5

43.6±8.8

MTE(Ours)

✓

X

18.7±2.2

14.2±1.3

44.8±2.2

MTE(Ours)

✓

X

19.7±1.6

15.8±1.3

44.3±1.8

Importance Weighting & MTL

with data augmentation

ZSL Model

Weighting Model

HMDB51

UCF101

Olympic Sports

RR [3]

–

Uniform

21.9±2.4

19.4±1.7

46.5±9.4

MTE (Ours)

Uniform

23.4±3.4

20.9±1.5

49.4±8.8

RR [3]

–

Visual KLIEP

23.2±2.7

20.3±1.6

47.2±9.3

RR [3]

–

Category KLIEP

23.6±2.1

20.2±1.6

51.8±1.7

RR [3]

–

Full KLIEP

23.7±2.7

20.7±1.4

51.3±0.9

MTE (Ours)

Visual KLIEP

23.4±2.8

20.8±2.0

51.4±2.9

MTE (Ours)

Category KLIEP

23.5±2.4

20.9±1.7

50.9±3.5

MTE (Ours)

Full KLIEP

23.9±3.0

21.7±2.7

52.3±1.8

Compare with the State-of-The-Art

With single-task models:

Method

Embed Feet TD

Aug HMDB51

UCF101

Olympic Sports

MTE

W

FV

X

19.7±1.6

15.8±1.3

44.3±8.8

MTE+Full KLIEP

W

FV

23.9±3.0

21.9±2.7

52.3±8.8

MTE+2 Full KLIEP

W

FV

24.8±3.2

22.8±3.3

56.4±7.7

MTE

A

FV

X

18.3±1.7

55.6±11.3

DAP [1] CVPR09

A

FV

X

15.9±1.2

45.4±12.8

IAP [1] CVPR09

A

FV

X

16.7±1.1

42.3±12.5

HA [6] CVPR11

A

FV

X

14.9±0.8

46.1±12.4

SVE [7] ICIP11

W

BoW

X

12.0±1.4

N/A

SVE [7] ICIP11

W

FV

X

18.4±1.4

N/A

ESZSL [8] ICL15

W

FV

X

15.0±1.3

39.6±9.6

ESZSL [8] ICL15

W

FV

X

14.9±1.2

39.5±10.8

SE [9] ICCV15

W

FV

X

17.5±2.4

28.6±4.9

SE [9] ICCV15

W

FV

X

12.0±1.7

47.5±14.8

Observations:

• Multi-task better than single-task model (RR).
• MTE model outperforms RMTL & GOMLT.
• NN matching in latent space is better

KLIEP Qualitative Study

Randomly selected 4 target videos, estimate the weight given to 16 auxiliary videos.

Qualitative Comparison: Single-Task vs. Multi-Task

Project testing videos into embedding space and apply T-SNE visualization for a 5 class ZSL exp.

Observations:

• MTE better than RR with the ROC curve is higher for each class.

References

Support Discrimination Dictionary Learning for Image Classification

Yang Liu¹, Wei Chen², Qingchao Chen³, Ian Wassell¹

¹ University of Cambridge, UK  ¨ Beijing Jiaotong University, China  ¤ University College London, UK

Dictionary Learning (DL) for Classification

- **Learn a class-specific dictionary**
  - Discriminates classes via representation residual.
  - **Limitation**
    - Sub-Dictionaries are disjoint to each other
    - How many and which atoms belong to each class is fixed
    - Not scalable to the problem with a large number of classes

- **Learn a shared dictionary**
  - A classifier is learned together by imposing some class-specific constraints on the coding vector.
  - **Limitation**
    - No multi-task setting has been used
    - Difficult to discriminate groups of coefficients owing to lack of prior knowledge of sub-dictionary structure

Support Discrimination DL Methodology

- **Design Principle**
  - Coefficients of images from the same class have a common sparse structure
  - Size of the overlapped signal support of different classes is minimized
  - Promote Intra-class similarity
  - Promote inter-class dissimilarity
  - Multi-task setting using a shared dictionary
  - Automatically identify overlapped sub-dictionaries
  - Size of each sub-dictionary is adjusted during learning
  - Scalable to allow for a large number of classes
  - Not using Euclidean distance to measure similarity
  - Achieves discrimination via support of representation
  - Structural sparse constraints ease the difficulty of solving the ill-posed inverse problem

Experiment

- **Factors Affecting Performance**
  - **Same Class:**
    - Common sparse structure
  - **Different Classes:**
    - Minimize overlapped support

- **Table 1. Recognition Rates (%) for Object Classification**

- **Table 2. Recognition Rates (%) for Face Classification**

Conclusion

- Multi-task setting using a shared dictionary
- Automatically identify overlapped sub-dictionaries
- Size of each sub-dictionary is adjusted during learning
- Scalable to allow for a large number of classes
- Not using Euclidean distance to measure similarity
- Achieves discrimination via support of representation
- Structural sparse constraints ease the difficulty of solving the ill-posed inverse problem

This work is funded by University of Cambridge Overseas Trust, EPSRC Research Grant (EP/K033700/1) AND Natural Science Foundation of China (61401018)
Accelerating the Super-Resolution Convolutional Neural Network
Chao Dong, Chen Change Loy, and Xiaou Tang
donchoa@setime.com  ccloy@ie.cuhk.edu.hk  xtag@ie.cuhk.edu.hk

**Motivation**

State of the art: Test on Set14

SRCNN (ECCV14) 0.39 s / 29.00 dB

SRCNN-Ex (PAMI15) 2.8 s / 29.30 dB

SCN (ICCV15) 3.6 s / 29.41 dB

ESPCN (CVPR 16) 0.12 s / 29.42 dB

Ours

FSRCNN 0.061 s / 29.43 dB

FSRCNN-s 0.023 s / 29.13 dB

**Fast Super-Resolution Conv. Neural Networks**

Introduce a deconvolution layer at the end to do upsampling.

Hourglass-shape, mapping => shrinking + mapping + expanding.

Smaller filter size, less filters, but deeper structure.

**SR for Different Upscaling Factors**

FSRCNN consists of convolution layers and a deconvolution layer. The convolution layers can be shared for different upscaling factors. A specific deconvolution layer is trained for different upscaling factors.

**Experiments**

The transitions from SRCNN to FSRCNN

<table>
<thead>
<tr>
<th>SRCNN-Ex</th>
<th>Transition 1</th>
<th>Transition 2</th>
<th>FSRCNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>First part</td>
<td>Conv(9,64,1)</td>
<td>Conv(9,64,1)</td>
<td>Conv(9,64,1)</td>
</tr>
<tr>
<td>Mid part</td>
<td>Conv(5,32,64)</td>
<td>Conv(5,32,64)</td>
<td>Conv(1,12,64)-4Conv(3,12,12)-Conv(1,64,12)</td>
</tr>
<tr>
<td>Last part</td>
<td>Conv(5,1,32)</td>
<td>DeConv(9,1,32)</td>
<td>DeConv(9,1,64)</td>
</tr>
</tbody>
</table>

Input size: HR LR LR LR

Parameters: 57184 58976 17088 12464

Speedup: 1× 8.7× 30.1× 41.3×

PSNR (Set5): 32.83 dB 32.95 dB 33.01 dB 33.06 dB

Towards real-time!

**Reference**


Symmetric Non-Rigid Structure from Motion for Category-Specific Object Structure Estimation
Yuan Gao¹ and Alan L. Yuille²,³
¹City University of Hong Kong, ²UCLA ³John Hopkins University

Motivation
Many objects, especially those that made by human, have intrinsic symmetry property, e.g., cars, acroplanes, etc.
Note that we assume the deformation between different objects is non-rigid and symmetric, e.g. from sedan to hatchback cars. While the object itself is rigid (also symmetric).

Our goal is to investigate how symmetry can improve NRSfM.

Contributions
By exploring symmetry, we proposed two NRSfM methods which gain significant improvement over their baseline counterparts.
Both of our methods can deal with occlusions, which are updated iteratively.

Sym-EM-PPCA: this is an extension of EM-PPCA [1]. It imposes symmetry constraints on both 3D structure and deformation bases. Sym-Rigid-SFM [5] is used to initialize Sym-EM-PPCA with hard symmetric constraints on the 3D structure.

Sym-PriorFree: it extends the matrix factorization methods of [2, 3] by symmetry, to initialize a coordinate descent algorithm for occlusions.

Ambiguities in NRSfM
The key idea of non-rigid SFM is to represent the non-rigid deformations of objects in terms of a linear combination of bases: \[ Y = RS \quad \text{and} \quad S = Vz. \]

\( RR^T = I \Rightarrow Y = RVz \quad \text{s.t.} \quad RR^T = I \)

Ambiguities 1: (between \( R \) and \( S \), will be solved by \( RR^T = I \))

\[ Y \rightarrow RA_1 \quad \text{and} \quad S \rightarrow A_1^{-1}S \]

Ambiguities 2: (between \( RV \) and \( z \), essentially “gauge freedom” as in [4])

\[ z \rightarrow A_2z \quad \text{and} \quad V \rightarrow VA_2^{-1}V \]

\[ z \rightarrow z + \alpha w \quad \text{and} \quad RV \rightarrow RV(z + \alpha w) \]

Different ways to deal with Ambiguities 2 leads to original EM-PPCA [1] and PriorFree [2,3] methods.

Sym-EM-PPCA method
It uses a conjugate (Gaussian) prior on \( z \) to deal with Ambiguities 2.
A mean shape is used to model the 3D structure of \( n \) th image:

\[ S_n = S + \mathbf{V}z_n \]

Let superscript \( \dagger \) denote symmetry, and \( \mathbf{Y}_n, \mathbf{S}, \mathbf{P}_n \) denote the stacked vectors of 2D keypoints, 3D mean structure, and translations. \( \mathbf{G}_n = I_p \otimes c_n \mathbf{R}_n \) in which \( c_n \) is the scale parameter, \( \mathbf{R}_n \) is the camera projection. \( \mathbf{V} \) is the deformation bases, \( z_n \) is the coefficient. After exploring symmetry, we have:

\[ P(Y_n | \mathbf{V}, \mathbf{S}, \mathbf{G}_n, \mathbf{T}_n) = N(V(g(S_n + \mathbf{V}z_n) + \mathbf{T}_n), \sigma^2) \]

Then, we maximize marginal probability \( P(Y_n | \mathbf{V}, \mathbf{G}_n, \mathbf{S}, \mathbf{V}, \mathbf{T}) \) by EM, where \( z_n \) is the latent variable with conjugate (Gaussian) prior.

E-step: maximize the posterior probability of \( z_n \):

\[ P(z_n | Y_n, \mathbf{V}, \mathbf{S}, \mathbf{G}_n, \mathbf{V}, \mathbf{T}) = N(V(g(S_n + \mathbf{V}z_n) + \mathbf{T}_n), \sigma^2) \]

M-step: maximize the joint likelihood with fixed \( z_n \):

\[ Q(\theta) = \sum_p \log P(Y_n | \mathbf{V}, \mathbf{S}, \mathbf{G}_n, \mathbf{V}, \mathbf{T}) + \log |V| - \lambda \| V - AV \|_F^2 \]

\[ = - \sum_p \log p(Y_n | \mathbf{V}, \mathbf{S}, \mathbf{G}_n, \mathbf{V}, \mathbf{T}) + \log |V| - \lambda \| V - AV \|_F^2 \]

Initialization: \( R_0, \mathbf{S}, \mathbf{T} \) and the occluded points \( \mathbf{Y}_{n:p}, \mathbf{Y}_{n:d} \) can be initialized by Bayesian PCA on the 2D KPs.

Sym-PriorFree method
As revealed in [4], Ambiguities 2 are in fact “gauge freedom”, which facilitates a prior free matrix factorization method [2,3].

The original prior free SFM model is [2,3]:

\[ Y = Rz \quad \text{and} \quad z = \sum_{r=1}^{n} r_z \]

\[ V = \Pi \]

where \( Y, R, S \) are stacked 2D keypoints, projection matrix and 3D structures. \( z \) is the coefficient associated with \( k \)‘th basis of image \( n \). \( V \) is the deformation bases, and \( \Pi = \mathbf{R} \odot I_z \).

The matrix factorization is conducted on \( Y \) w.r.t \( \Pi \) and \( V \):

\[ \min \| Y - \Pi V \|_F^2 \quad \text{(because rank}(Y) = \min(2N, 3K, P)) \]

After imposing symmetry, we have: (superscript \( \dagger \) denotes symmetry)

\[ Q(R, S) = \| Y - RS \|_F^2 + \| Y' - RS' \|_F^2 = \| Y - \Pi V \|_F^2 + \| Y' - \Pi V' \|_F^2 \]

Note that we cannot do SVD on \( Y \) and \( Y' \) separately, because the two energies are dependent. In other words, doing SVD separately cannot ensure to provide the same \( \Pi \) and the symmetric \( V, V' \).

Therefore, we further explore symmetry to decouple the dependency. Assume the object is symmetric along x-axis in world coordinate system (we can always rotate the world coordinate system to ensure that), we have:

\[ Y = \Pi V = \Pi \Pi^* \]

\[ Y' = \Pi \Pi' \]

Therefore, the decoupled energies are:

\[ Q(\Pi, V) = \| Y - \Pi V \|_F^2 + \| M - \Pi \Pi^* V \|_F^2 \]

The SVD (initial) on \( L \) and \( M \) gives us \( \Pi_1, V_1, \Pi_2, V_{1g} \), which is up to ambiguities \( H^p, H^f \) to the true \( \Pi_1, \Pi_2, V_1, V_{1g} \).

\[ L = \Pi_1 V_1 = \Pi_1 H^p (H^f)^{-1} V_1 \quad \text{and} \quad M = \Pi_2 V_{1g} = \Pi_2 H^f (H^p)^{-1} V_{1g} \]

These ambiguities can be solved by applying orthonormality constraints \( R^T = I, \odot \Pi \), where \( \Pi = \Pi_1^*, \Pi_2^* \). The (under-determined) solutions lie in the intersection of three subspaces \([2,3], \text{i.e. any solution that satisfies these three subspaces gives the same 3D reconstruction results, only up to different rotations and scales [4]. Please refer to our paper for the details for solving } H^p \text{ and } H^f \)

After solving \( H^p \) and \( H^f \), we can recover \( R \) by normalizing each row of \( \Pi \) to have unit L2 norm, then \( S \) can be estimated accordingly by minimizing the reconstruction error with low-rank prior on \( S \).

Results
Please see our paper for more results! (The last two rows are ours)
Pascal3D+ dataset [6] is used. The Roman numerals are the mean shape error for each subtypes (i.e. sedan car, because we only have groundtruth shape for subtypes), mRE is the mean rotation error for each category.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>E</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>X</th>
<th>N</th>
<th>mRE</th>
<th>mRE</th>
<th>nre</th>
<th>nre</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>1.10</td>
<td>1.01</td>
<td>1.00</td>
<td>1.00</td>
<td>1.03</td>
<td>1.07</td>
<td>1.09</td>
<td>1.94</td>
<td>1.38</td>
<td>0.58</td>
<td>0.57</td>
<td>0.47</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>IP+</td>
<td>1.76</td>
<td>1.67</td>
<td>1.57</td>
<td>1.65</td>
<td>1.76</td>
<td>1.89</td>
<td>1.97</td>
<td>2.10</td>
<td>1.55</td>
<td>1.43</td>
<td>1.29</td>
<td>1.06</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

We also investigate the noise annotations, as it affect the symmetry. Experiments on aeroplane with different Gaussian noise \( \sigma \) show that our methods is robust to noise.

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>E</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
<th>VII</th>
<th>VIII</th>
<th>IX</th>
<th>X</th>
<th>N</th>
<th>mRE</th>
<th>mRE</th>
<th>nre</th>
<th>nre</th>
</tr>
</thead>
<tbody>
<tr>
<td>IP</td>
<td>0.31</td>
<td>0.30</td>
<td>0.40</td>
<td>0.41</td>
<td>0.34</td>
<td>0.45</td>
<td>0.43</td>
<td>0.33</td>
<td>0.37</td>
<td>0.38</td>
<td>0.50</td>
<td>0.47</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>IP+</td>
<td>0.24</td>
<td>0.34</td>
<td>0.47</td>
<td>0.44</td>
<td>0.22</td>
<td>0.60</td>
<td>0.52</td>
<td>0.32</td>
<td>0.33</td>
<td>0.54</td>
<td>0.47</td>
<td>0.43</td>
<td>0.43</td>
<td></td>
</tr>
</tbody>
</table>

Reference
Peak-Piloted Deep Network for Facial Expression Recognition

Xiangyun Zhao\textsuperscript{1} Xiaodan Liang\textsuperscript{2} Luoqi Liu\textsuperscript{3} Teng Li\textsuperscript{4} Yugang Han\textsuperscript{3} Nuno Vasconcelos\textsuperscript{1} Shuicheng Yan\textsuperscript{3}

\textsuperscript{1}Statistical Visual Computing Laboratory, University of California, San Diego  \textsuperscript{2}Carnegie Mellon University  \textsuperscript{3}360 AI Institute  \textsuperscript{4}Institute of Automation, Chinese Academy of Sciences

1. Introduction

In this work, we present a novel peak-piloted deep network (PPDN) that uses a sample with peak expression (easy sample) to supervise the intermediate feature responses for a sample of non-peak expression (hard sample) of the same type and from the same subject. The expression evolving process from non-peak expression to peak expression can thus be implicitly embedded in the network to achieve the invariance to expression intensities.

Fig1. Expression evolving process from non-peak expression to peak expression

2. Motivation

It is usually difficult to capture critical and subtle expression details from non-peak expression images, which can be hard to distinguish across expressions. In principle, an mapping from non-peak to peak expression could improve recognition.

Fig2. Examples of six facial expression samples, including surprise, angry, happy, fear, sad and disgust. For each subject, the peak and non-peak expressions are shown

3.1 Network Structure

During training, PPDN takes the pair of peak and non-peak expression images as input. After passing the pair through several convolutional and fully-connected layers, the intermediate feature maps can be obtained for peak and nonpeak expression images, respectively. The L2-norm loss between these feature maps is optimized for driving the features of the non-peak expression image towards those of the peak expression image. The network parameters can thus be updated by jointly optimizing the L2-norm losses and the losses of recognizing two expression images. During the back-propagation process, the Peak Gradient Suppression (PGS) is utilized.

3.2 Network Optimization

We denote the training set as \( S = \{ x^p_i, y^p_i, x^n_i, y^n_i, i = 1 ... N \} \), where \( x^p_i \) denotes a face with non-peak expression, \( x^n_i \) a face with corresponding peak expression, and \( y^p_i \) and \( y^n_i \) are corresponding expression labels. Let J1, J2 and J3 indicate the L2-norm of the feature differences and the two cross-entropy losses for recognition, respectively. The proposed peak gradient suppression (PGS) learning algorithm uses instead the updates

\[
W^{\text{new}} = W - \gamma \left( \frac{\partial J_1}{\partial W} x^n_i y^n_i + \frac{\partial J_2}{\partial W} x^n_i y^n_i \right) \frac{\partial J_1}{\partial W} x^n_i y^n_i - \frac{1}{N} \sum_{i=1}^{N} \frac{\partial J_1}{\partial W} x^n_i y^n_i - 2 W
\]

Where \( \gamma \) is learning rate. The feature responses of the peak expression image are suppressed in PGS.

4. Experiments

To evaluate the PPDN, we conduct extensive experiments on two popular FER datasets: CK+ \cite{1} and Oulu-CASIA \cite{2}. To further demonstrate that the PPDN generalizes to other recognition tasks, we also evaluate its performance on face recognition over the public Multi-PIE dataset \cite{3}.

Table1. Performance comparisons on six facial expressions with four state-of-the-art methods and the baseline on CK+ database.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSFR \cite{10}</td>
<td>86.9%</td>
</tr>
<tr>
<td>AdaCohor \cite{14}</td>
<td>93.3%</td>
</tr>
<tr>
<td>LBPSVM \cite{11}</td>
<td>95.1%</td>
</tr>
<tr>
<td>BDBN \cite{4}</td>
<td>96.7%</td>
</tr>
<tr>
<td>GoogleLeNet(baseline)</td>
<td>95.0%</td>
</tr>
<tr>
<td>PPDN</td>
<td>97.3%</td>
</tr>
</tbody>
</table>

Table2. Performance comparisons on six facial expressions with four state-of-the-art methods and the baseline on Oulu-CASIA database.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDCS \cite{5}</td>
<td>49.9%</td>
</tr>
<tr>
<td>GoogleLeNet(baseline)</td>
<td>66.6%</td>
</tr>
<tr>
<td>PPDN</td>
<td>72.4%</td>
</tr>
</tbody>
</table>

Table3. Performance comparison on CK+ when evaluating on three different test sets, including “weak expression”, “peak expression” and “combined”, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>weak expression</th>
<th>peak expression</th>
<th>combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDN(standard SGD)</td>
<td>81.34%</td>
<td>99.12%</td>
<td>94.18%</td>
</tr>
<tr>
<td>GoogleLeNet(baseline)</td>
<td>78.10%</td>
<td>98.96%</td>
<td>92.19%</td>
</tr>
<tr>
<td>PPDN</td>
<td>83.36%</td>
<td>99.30%</td>
<td>95.33%</td>
</tr>
</tbody>
</table>

Table4. Performance comparison on Oulu-CASIA database when evaluating on three different test sets, including “weak expression”, “peak expression” and “combined”, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>weak expression</th>
<th>peak expression</th>
<th>combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPDN(standard SGD)</td>
<td>67.05%</td>
<td>82.91%</td>
<td>73.54%</td>
</tr>
<tr>
<td>GoogleLeNet(baseline)</td>
<td>64.64%</td>
<td>79.21%</td>
<td>71.32%</td>
</tr>
<tr>
<td>PPDN</td>
<td>67.95%</td>
<td>84.59%</td>
<td>74.99%</td>
</tr>
</tbody>
</table>

Table5. Face recognition rates for various poses on Multi-PIE with some state of the art.

<table>
<thead>
<tr>
<th>Method</th>
<th>(-45\degree)</th>
<th>(-30\degree)</th>
<th>(-15\degree)</th>
<th>(+15\degree)</th>
<th>(+30\degree)</th>
<th>(+45\degree)</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Li et al. \cite{38}</td>
<td>56.62%</td>
<td>77.22%</td>
<td>89.11%</td>
<td>88.81%</td>
<td>79.12%</td>
<td>58.14%</td>
<td>78.48%</td>
</tr>
<tr>
<td>Zhu et al. \cite{27}</td>
<td>61.00%</td>
<td>74.60%</td>
<td>86.10%</td>
<td>83.30%</td>
<td>75.30%</td>
<td>61.80%</td>
<td>74.70%</td>
</tr>
<tr>
<td>CPI \cite{28}</td>
<td>66.60%</td>
<td>78.00%</td>
<td>87.30%</td>
<td>83.30%</td>
<td>75.30%</td>
<td>62.30%</td>
<td>75.90%</td>
</tr>
<tr>
<td>CPI \cite{28}</td>
<td>73.00%</td>
<td>81.70%</td>
<td>89.40%</td>
<td>89.90%</td>
<td>80.50%</td>
<td>70.30%</td>
<td>80.70%</td>
</tr>
<tr>
<td>GoogleLeNet(baseline)</td>
<td>56.62%</td>
<td>77.22%</td>
<td>89.11%</td>
<td>88.81%</td>
<td>79.12%</td>
<td>58.14%</td>
<td>74.84%</td>
</tr>
<tr>
<td>PPDN</td>
<td>72.06%</td>
<td>85.41%</td>
<td>92.14%</td>
<td>91.38%</td>
<td>87.07%</td>
<td>76.97%</td>
<td>83.21%</td>
</tr>
</tbody>
</table>


\[4\] SVCL
Is Faster R-CNN Doing Well for Pedestrian Detection?
Liliang Zhang1, Liang Lin1, Xiaodan Liang1, Kaiming He2
1Sun Yat-sen University 2Microsoft Research
(Source code: https://github.com/zhangliliang/RPN_BF)

Challenges for Faster/Fast R-CNN
- Small objects that may fail RoI pooling on low-resolution feature maps.
- Hard negative examples that receive no careful attention.

Our Pipeline
- RPN is used to compute candidate bounding boxes, scores, and convolutional feature maps.
- The candidate boxes are fed into cascaded Boosted Forests (BF) for classification, using the features pooled from the convolutional feature maps computed by RPN.

Takeaway messages
- RPN is good for pedestrian detection.
- Feature resolution is very important.
- Bootstrapping/Hard negative mining helps.

State-of-the-art Performance
- Caltech (IoU > 0.5)
- Caltech (IoU > 0.7)
- Caltech-New
- INRIA
- ETH
- KITTI

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP on Easy</th>
<th>mAP on Moderate</th>
<th>mAP on Hard</th>
<th>Times (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-CNN</td>
<td>61.61</td>
<td>50.13</td>
<td>44.79</td>
<td>4</td>
</tr>
<tr>
<td>pAUCFastT</td>
<td>65.26</td>
<td>54.49</td>
<td>48.69</td>
<td>60</td>
</tr>
<tr>
<td>FilteredHCF</td>
<td>67.65</td>
<td>56.75</td>
<td>51.12</td>
<td>2</td>
</tr>
<tr>
<td>DeepPart</td>
<td>70.49</td>
<td>58.67</td>
<td>52.78</td>
<td>1</td>
</tr>
<tr>
<td>CompACT-Deep</td>
<td>70.69</td>
<td>58.74</td>
<td>52.71</td>
<td>1</td>
</tr>
<tr>
<td>Regolets</td>
<td>73.14</td>
<td>61.15</td>
<td>55.21</td>
<td>1</td>
</tr>
<tr>
<td>RPN+BFR</td>
<td>77.12</td>
<td>61.15</td>
<td>55.12</td>
<td>0.6</td>
</tr>
</tbody>
</table>

(Source code: https://github.com/zhangliliang/RPN_BF)
1 Introduction

Simultaneous localization and mapping (SLAM) using the whole image data is an appealing framework to address shortcoming of sparse feature-based methods – in particular frequent failures in textureless environments. Hence, direct methods bypassing the need of feature extraction and matching became recently popular. Many of these methods operate by alternating between pose estimation and computing (semi-)dense depth maps, and are therefore not fully exploiting the advantages of joint optimization with respect to depth and pose.

Convex-to-fine Planar Regularization

for Dense Monocular Depth Estimation

Stephan Liwicki
Christopher Zach

3 Optimization Strategy

We find pose as well as depth through Levenberg-Marquardt optimization.

- Locally linearize $f_p$ (as in KLT)
- Graduated optimization [3]
- scale-space pyramid
- proposed restricted depth

Figure 3: We incrementally increase depth resolution.

Algorithm 1: Dense Incremental Planar Depth Estimation

$\text{Coarse-to-fine Depth Perception}$

We enforce low complexity in the update of $S$:

$$s_i = s^*_i + \sum I(x_i) \Delta c_i,$$

$\text{Advantages in Optimization Conditions}$

- Regularization across all pixels is enforced for the update of planes
- The image planes encode inverse depth hierarchically from coarse to fine
- The incremental algorithm enables fast, joint optimization of pose and depth
- Depth updates are simple, but yield rich depth maps after repeated updates

4 Evaluation

We evaluate our methods, dense DIP and semi-dense SIP, on 6 TUM and 7 own videos.

$\text{Quantitative Results}$

- DIP (one keyframe) performs similar to LSD-SLAM (many keyframes) for pose
- DIP benefits from larger baselines for depth estimation
- LSD-Key (LSD-SLAM with single reference frame) is less favourable
- Our semi-dense SIP performs well for small motion
- The Disjoint (alternating) version is worse in virtually all experiments

$\text{Qualitative Results}$

- SIP, DIP produce more globally consistent depths in contrast to LSD-SLAM
- Even in non-planar scenes, our method produces sensible results
- Our method converges to local minima if initial $s_i \leftarrow [0.01]^{12}$ is significantly wrong

$\text{Running time}$

- DIP is highly parallel and computes in real-time on GPU
- SIP estimates depth and pose twice as fast as LSD-SLAM using CPU (30 fps)
Deep Attributes Driven Person Re-identification

Chi Su1, Shiliang Zhang1, Junliang Xing2, Wen Gao1, and Qi Tian3

Introduction

- **Motivation:** Attribute: a stable and robust mid-level feature, we try to extract attributes as features.
  - Deep Convolutional Neural Network (dCNN) has very powerful learning ability, good performance in pattern recognition tasks and good generalization ability.
- **Methodology:** We propose a semi-supervised deep attribute learning (SSDAL) framework which progressively boosts the accuracy of attributes only using a limited number of labeled data.

Stage 1: Fully-supervised dCNN training

- Independent dataset with attribute labels

Stage 2: Fine-tuning using attributes triplet loss

- Dataset with person ID labels
- Predicted attributes

Stage 3: Final fine-tuning on the combined dataset

- Improved dCNN predicts attributes labels for the target dataset, which is finally combined with the independent dataset for the second round of fine-tuning.

Examples of Predicted Attributes:

- **Training dataset:**
  - PETA: annotated with 105 attributes.
  - MOT challenge: annotated with person ID.
- **Independent testing dataset:**
  - VIPeR: 632 persons, each has two 48*128 images.
  - PRID: 385 and 749 persons captured by camera A and B.
  - Market: 25,000 images of 1501 persons taken by 6 cameras.

Experiments

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank 1</th>
<th>Rank 5</th>
<th>Rank 10</th>
<th>Rank 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>VIPeR</td>
<td>4.9</td>
<td>14.0</td>
<td>23.6</td>
<td>39.2</td>
</tr>
<tr>
<td>PETA</td>
<td>4.5</td>
<td>12.4</td>
<td>20.7</td>
<td>28.7</td>
</tr>
<tr>
<td>MOT</td>
<td>4.5</td>
<td>12.7</td>
<td>20.7</td>
<td>28.7</td>
</tr>
<tr>
<td>Shangai</td>
<td>5.5</td>
<td>13.3</td>
<td>22.3</td>
<td>37.1</td>
</tr>
<tr>
<td>JCDA</td>
<td>5.5</td>
<td>13.3</td>
<td>22.3</td>
<td>37.1</td>
</tr>
<tr>
<td>Galleria</td>
<td>4.5</td>
<td>12.2</td>
<td>21.3</td>
<td>35.2</td>
</tr>
<tr>
<td>LMDA</td>
<td>5.0</td>
<td>18.3</td>
<td>38.1</td>
<td>66.6</td>
</tr>
<tr>
<td>LMDA+XID</td>
<td>5.0</td>
<td>18.3</td>
<td>38.1</td>
<td>66.6</td>
</tr>
<tr>
<td>LMDA+XID+</td>
<td>5.0</td>
<td>18.3</td>
<td>38.1</td>
<td>66.6</td>
</tr>
</tbody>
</table>

Acknowledgement

This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0290 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar. This work was supported in part by National Science Foundation of China (NSFC) 61422007 and 61302178. This work was supported in part to Dr. Shiliang Zhang by National Science Foundation of China (NSFC) 61572050 and 9153811.

Contact Information

Chi Su (chisu@pku.edu.cn)
Shiliang Zhang (slzhang.jd@pku.edu.cn)
An Occlusion-Resistant Ellipse Detection Method by Joining Coelliptic Arcs

Halil Ibrahim Cakir(1), Cihan Topal(2,3), Cuneyt Akinlar(4)

(1) Computer Engineering, Dumlupinar University, Turkey
(2) Electrical and Electronic Engineering, Anadolu University, Turkey
(3) Visea Innovative, Eskisehir, Turkey
(4) Computer Engineering, Anadolu University, Turkey

1. Ellipse Detection

- An active research problem due to its difficulty.
- The projections of circular objects appear as ellipse on the camera image plane.
- Occlusions in real life images significantly aggravates the problem.
- Fast and robust ellipse detection is a vital step in many computer vision applications.

2. Proposed Method

Algorithm:
1. Detection of edge segments (Akinlar and Topal, 2012)
2. Computing corners by a CSS based algorithm (Topal et al. 2013)
3. Extracting elliptical arcs by fitting ellipse to the pixels between two consecutive corners
4. Arc pairing strategy is deployed by using conic properties of arcs
5. Ellipse fitting is applied to remaining arc combinations to decide on final ellipses.

2.1 Arc Detection

(i) Fitting an ellipse to pixel chains lying between two consecutive corners.
(ii) Computation of fitting error to decide whether it is an elliptical arc or not.

2.2 Detection of Non-Coelliptic Arc Pairs

Problem:
(i) \( 2^n - 1 \) potential combinations to join and detect ellipses \( (2^{10} - 1 = 65535) \) for test image.
(ii) Ellipse fitting and error computation are also computationally expensive.

Solution:
- Analyzing all arcs in pairwise to determine whether any pair can be part of the same ellipse or not.
- Find out whether any two arcs are coelliptic or non-coelliptic
- Complexity of arc pairing operation is polynomial \( (O(n^2)) \) instead of \( (O(2^n)) \)

2.1.1 Arc Pairing Test:
\[ A_x x^2 + B_x y + C_y y^2 + D_y x + E_y y + F_y = 0 \]  \( \text{(1)} \)
\[ A_x x + B_y y + F_y = 0 \]  \( \text{(2)} \)
Find the \( m \) and \( n \) values in the line equation:
\[ m = \frac{y_2 - y_1}{x_2 - x_1} \]
\[ n = y_1 - mx_1 \]
Substitute \( mx + n \) in place of \( y \) in the conic equation:
\[ x^2(A_1 + mx + nC_1) + x(nB_1 + 2mA_1 + D_1 + nE_1) + (C_1n^2 + E_1n + F_1) = 0 \]  \( \text{(5)} \)
Solve it with discriminant analysis:
- Analyzing the intersection point actually lies on the arc.

Arc pairing table for test image:

<table>
<thead>
<tr>
<th>Arc 1</th>
<th>Arc 2</th>
<th>Pairing</th>
<th>Coelliptic</th>
<th>Non-Coelliptic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A2</td>
<td>A1 A2</td>
<td>C</td>
<td>NC</td>
</tr>
<tr>
<td>A2</td>
<td>A1</td>
<td>A2 A1</td>
<td>C</td>
<td>NC</td>
</tr>
<tr>
<td>A3</td>
<td>A4</td>
<td>A3 A4</td>
<td>C</td>
<td>NC</td>
</tr>
<tr>
<td>A4</td>
<td>A3</td>
<td>A4 A3</td>
<td>C</td>
<td>NC</td>
</tr>
</tbody>
</table>

“C” and “NC” stand for coelliptic and non-coelliptic, and indicates that the arc pair can be joined or not, respectively.

2.3 Detection of Final Ellipses

(i) RMS error is small, i.e. \( \leq 2 \) pixels
(ii) SAR is around 0.5 w.r.t. eccentricity

Notice that some cases (last column) may violate our heuristic although it is valid in general.

3. Experimental Results

- Three annotated real datasets containing 1227 images (Fornaciari, 2014; Prasad, 2012).
- Intel i7 2.40 GHz processor with 8 GB of RAM
- Both saves excessive computation time, and handles occluded situations.
- Obtain better accuracy results compare to existing methods in the literature.
- Gives also promising results in terms of computation time.

<table>
<thead>
<tr>
<th>Dataset / Algorithm</th>
<th>Prasad</th>
<th>Libuda</th>
<th>Fornaciari</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone Dataset</td>
<td>2.97 ms</td>
<td>5.5 ms</td>
<td>5.6 ms</td>
<td>1.0 ms</td>
</tr>
<tr>
<td>Random Dataset</td>
<td>1.842 ms</td>
<td>9.5 ms</td>
<td>9.3 ms</td>
<td>3.88 ms</td>
</tr>
<tr>
<td>Prasad Dataset</td>
<td>1.780 ms</td>
<td>27.2 ms</td>
<td>14.9 ms</td>
<td>16.8 ms</td>
</tr>
</tbody>
</table>
Branching Path Following for Graph Matching

Tao Wang 1,2, Haibin Ling 1,3, Congyan Lang 1, Jun Wu 2

1Meitu HiScene Lab. HiScene Information Technologies, Shanghai, China
2School of Computer & Information Technology, Beijing Jiaotong University, Beijing 100044, China
3Computer & Information Sciences Department, Temple University, Philadelphia 19122, USA

Email: twang@bjtu.edu.cn, hbling@temple.edu, cylang@bjtu.edu.cn, wuj@bjtu.edu.cn

1. Introduction

- Existing path following algorithms for graph matching can be viewed as special cases of the numerical continuation method (NCM), and correspond to particular implementation named generic predictor corrector (GPC).
- The GPC approach succeeds at regular points, but may fail at singular points. Illustration of GPC and the proposed method is shown in Fig. 1.
- This paper presents a branching path following (BPF) method to exploring potentially better paths at singular points to improve matching performance.

2. Proposed Method

2.1 Singular Points Discovery

- The path following strategy results in a constrained nonlinear system \( F(\lambda, u) = 0 \), s.t. \( \lambda_0 \).
- Denote \( J \) the Jacobian of \( F \) parameterized by \( \lambda \). A singular point \( (\lambda, u) \) should have \( |J_u| = 0 \).
- A reasonable assumption: the curve formed by \( (\lambda, u(\lambda)) \) over \( \lambda \) is continuous. We thus mark \( (\lambda, u) \) as a singular point if \( |J_u| |J_{\lambda u}| \leq 0 \).
- Signs of determinants of Jacobians are computed via LU decomposition.

2.2 Branch Switching

- Using the pseudo-arclength continuation algorithm to explore a new branch at each singular point:
  \[
  G(\lambda, u) = \begin{bmatrix}
  F(\lambda, u) \\
  N(\lambda, u, s)
  \end{bmatrix} = \begin{bmatrix}
  0 \\
  0
  \end{bmatrix}
  \]
  \[
  N(\lambda, u, s) = (u')^T (u - u_0) + (\lambda')^T (\lambda - \lambda_0) - s = 0
  \]
  where \( (\lambda', u') \) denotes the tangent vector at point \( (\lambda, u) \), approximated using previous points as
  \[
  (\lambda', u') = (\Delta \lambda, \sum_{i=1}^{k-1} (k-1)(u_{i+1} - u_i))
  \]
- We solve \( G(\lambda, u) = 0 \) using the trust-region-reflective algorithm.

2.3 Upgrading existing algorithms

- The proposed BPF methods are ready to be combined with algorithms that utilize the path following strategy;
- We integrate BPF into a state-of-the-art graph matching algorithms, GNCCP, denote the enhanced algorithms BPF-G.

3. Experiments

3.1 Synthetic Dataset

- Three different settings varying the number of outliers, edge density and edge noise respectively. 100 pairs of random graphs for each setting.
- The proposed BPF-G algorithm achieves the best performance in terms of both objective ratio and matching accuracy.

3.2 Real Image Datasets

- Four real image datasets: CMU House, Pascal, Willow and Caltech.
- The proposed BPF-G algorithm achieves the best or nearly the best performance in all these datasets.

4. Conclusion

- We propose an efficient method of singular point discovery in the path of the GPC approach, and drive an branch switching to explore potentially better paths at singular points;
- Our approach offers remarkable improvement in both objective ratio and matching accuracy.
**Introduction**

- Semantic Segmentation is the task of labelling every pixel with its object category.
- Fully convolutional networks (FCNs) classify pixels independently of each other, and produce noisy predictions which do not respect image edges.
- As a result, Conditional Random Fields (CRF) with pairwise terms perform spatial and appearance consistency, are usually used as post-processing.
- We formulate a richer and more expressive CRF model which utilises Higher Order Potentials (potentials defined over cliques of more than two variables).
- We use the differentiable Mean Field inference algorithm to obtain the most probable labelling, and incorporate it as a layer of our neural network.
- This allows us to train our Higher Order CRF end-to-end with an FCN.

**Higher Order Potentials**

- We introduce two types of higher order potential into our differentiable CRF:
  - Detection potential uses the complementary cues of an object detector to improve segmentations. It helps in cases where initial unaries are poor.
  - Superpixel potential encourages consistency over larger regions, and removes spurious noise from the output.

**Formulation**

A Conditional Random Field is defined as

\[
P_t(X = x | I) = \frac{1}{Z(I)} \exp(-E(x | I)).
\]

In our case, the energy (ignoring conditioning on Image I) is:

\[
E(x) = \sum_{\text{Unaries from CNN}} v_i(x_i) + \sum_{\text{Pairwise}} v_{ij}(x_i, x_j) + \sum_{\text{Detection potentials}} v_{i}(x_i) + \sum_{\text{Superpixel potentials}} v_{ij}(x_{ij})
\]

**Detection Potential**

Our detection uses the output of an object detector as additional cues for segmentation. Intuitively, object detectors can help when our pixelwise predictions are incorrect.

- Assume we have D object detections for a given image.
- The \(d^{th}\) detection is of the form \((x_d, s_d, F_d)\):
  - \(L_d \in \mathcal{L}\) is the class label of the \(d^{th}\) detection.
  - \(s_d\) is the detection score. detection.
  - \(F_d \subseteq \{1, 2, \ldots, N\}\) is the set of pixels belonging to the detection foreground, obtained using GrabCut.
- Introduce binary latent variables, \(y_1, y_2, \ldots, y_N\) — one for each detection.
- Models whether detection is accepted or not.
- \(P(y_d = 1)\) initialised with \(s_d\), the score of the object detector.
- \(w_{d,\text{lin}}(x_d)\) is a learnable weight parameter that is a function of the class label.

This potential encourages consistency between detections, \(Y\), and labelled pixels, \(X\):

\[
\psi^D_{d}(X_d = x_d, Y_d = y_d) = \begin{cases} 
1 & \text{if } y_d = 0 \\
\exp \left( - \sum_{x_j \in F_d} w_{d,\text{lin}}(x_d) \right) & \text{if } y_d = 1 
\end{cases}
\]

**Superpixel Potential**

Our learnable superpixel potential enforces consistency over regions obtained by superpixels. This is a soft constraint using a \(P^D\)-Potts type energy. We use superpixels over multiple scales, which do not necessarily have to form a hierarchy.

\[
\psi^S_{ij}(X_{ij} = x_{ij}) = \frac{m_{ij}^\text{low}}{m_{ij}^\text{high}} \quad \text{if all } x_{ij}^l = l \\
\text{otherwise.}
\]

**Experimental Results on PASCAL VOC**

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean IoU (%)</th>
<th>Method</th>
<th>Mean IoU (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>77.9</td>
<td>Baseline (Unary + Pairwise)</td>
<td>72.9</td>
</tr>
<tr>
<td>Dilated [7]</td>
<td>75.3</td>
<td>BoxSup [8]</td>
<td>75.2</td>
</tr>
</tbody>
</table>

**Conclusion**

- Introduced two higher order potentials for a CRF which can be integrated into a deep neural network and trained end-to-end.
- Achieved the best performance on the PASCAL VOC dataset.
- In subsequent work [10], we have showed how our Detection potential can be used for the task of Instance Segmentation.
LSTM-CF: Unifying Context Modeling and Fusion with LSTMs for RGB-D Scene Labeling

Zhen Li1  Yukang Gan2  Xiaodan Liang2  Yizhou Yu1  Hui Cheng2  Liang Lin2
The University of Hong Kong1  Sun Yat-sen University2

Motivations

- Indoor scene labeling is more difficult than outdoor scene labeling
- Spatial layout of a scene and the relationship between different objects are important cues
- Discriminative CNN features of RGB and depth channels
- Fusion of depth and photometric data channels
- Long-range dependencies in scene labeling

Advantages

- Explicit context modeling through two parallel memorized context layers
- Fine-tuning DeepLab for color channels
- Simplified CNN layers for encoded depth channels
- A memorized fusion layer to extract true 2D context in a data-driven way
- Multi-scale features and cross-layer combination to enhance feature representation

A New Pipeline for RGB-D Scene Labeling

Refining Newly Labeled Ground Truth

As LSTM-CF is an end-to-end context model, it can recognize areas with occlusions through context inference. Thus, our model can be exploited to refine the newly labeled ground truth in the SUNRGBD dataset.

Contributions

- Our LSTM-CF model captures image contexts from a global perspective and deeply fuses contextual information from multiple sources.
- Our LSTM-CF model jointly optimizes LSTM layers and convolutional layers.
- Context modeling and fusion are effectively incorporated into the deep network architecture.
- Our model achieves the state-of-the-art performance on SUNRGBD, NYU-depth v2. It can be leveraged to improve the groundtruth annotations of newly captured 3943 RGB-D images in the SUNRGBD dataset.

Acknowledgements

Joint work supported by Projects on Faculty/Student Exchange and Collaboration Scheme between the Higher Education in Hong Kong and the Mainland, Guangzhou Science and Technology Program under grant 1563600439, and Fundamental Research Funds for the Central Universities and also supported by Hong Kong Research Grant Council under General Research Fund. Zhen Li is supported by Lee Shau Kee Postgraduate Fellowship from the University of Hong Kong.

Experiments on SUNRGBD Dataset

Class and Mean accuracy

<table>
<thead>
<tr>
<th>Method</th>
<th>Scene</th>
<th>Wall</th>
<th>Floor</th>
<th>door</th>
<th>window</th>
<th>ceiling</th>
<th>table</th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>lamp</th>
<th>tv</th>
<th>sink</th>
<th>bathhub</th>
<th>box</th>
<th>fridge</th>
<th>blinds</th>
<th>curtain</th>
<th>nightstand</th>
<th>picture</th>
<th>4.1</th>
<th>14.8</th>
<th>23.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Song2015</td>
<td>37.8</td>
<td>45</td>
<td>7.1</td>
<td>2.7</td>
<td>1</td>
<td>119.1</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu2011</td>
<td>38.9</td>
<td>43.2</td>
<td>18.8</td>
<td>3.2</td>
<td>4</td>
<td>12.9</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin2015</td>
<td>37.5</td>
<td>46.5</td>
<td>17.2</td>
<td>22.7</td>
<td>21.2</td>
<td>21.2</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Song2015</td>
<td>37.5</td>
<td>46.5</td>
<td>17.2</td>
<td>22.7</td>
<td>21.2</td>
<td>21.2</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu2011</td>
<td>38.9</td>
<td>43.2</td>
<td>18.8</td>
<td>3.2</td>
<td>4</td>
<td>12.9</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin2015</td>
<td>37.5</td>
<td>46.5</td>
<td>17.2</td>
<td>22.7</td>
<td>21.2</td>
<td>21.2</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ablation study

- 2.2% improvement over previous state-of-the-art average accuracy and best on 15 categories
- Photometric data is vital for scene labeling while depth is an auxiliary information source.
- LSST-CF model is important for fusing context features from different sources.

Experiments on NYU-Depth v2 and SUN3D Datasets

Class and Mean accuracy for NYU-Depth v2

<table>
<thead>
<tr>
<th>Method</th>
<th>Scene</th>
<th>Wall</th>
<th>Floor</th>
<th>door</th>
<th>window</th>
<th>ceiling</th>
<th>table</th>
<th>chair</th>
<th>toilet</th>
<th>bed</th>
<th>lamp</th>
<th>tv</th>
<th>sink</th>
<th>bathhub</th>
<th>box</th>
<th>fridge</th>
<th>blinds</th>
<th>curtain</th>
<th>nightstand</th>
<th>picture</th>
<th>4.1</th>
<th>14.8</th>
<th>23.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Song2015</td>
<td>37.8</td>
<td>45</td>
<td>7.1</td>
<td>2.7</td>
<td>1</td>
<td>119.1</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu2011</td>
<td>38.9</td>
<td>43.2</td>
<td>18.8</td>
<td>3.2</td>
<td>4</td>
<td>12.9</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin2015</td>
<td>37.5</td>
<td>46.5</td>
<td>17.2</td>
<td>22.7</td>
<td>21.2</td>
<td>21.2</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Song2015</td>
<td>37.8</td>
<td>45</td>
<td>7.1</td>
<td>2.7</td>
<td>1</td>
<td>119.1</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liu2011</td>
<td>38.9</td>
<td>43.2</td>
<td>18.8</td>
<td>3.2</td>
<td>4</td>
<td>12.9</td>
<td>74.4</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin2015</td>
<td>37.5</td>
<td>46.5</td>
<td>17.2</td>
<td>22.7</td>
<td>21.2</td>
<td>21.2</td>
<td>60.4</td>
<td>104.4</td>
<td>41.1</td>
<td>33.6</td>
<td>45.00</td>
<td>7.1</td>
<td>6.5</td>
<td>73.2</td>
<td>5.5</td>
<td>1.4</td>
<td>5.7</td>
<td>23.3</td>
<td>56.4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ablation study

- 2.2% improvement over previous state-of-the-art average accuracy and best on 15 categories
- Photometric data is vital for scene labeling while depth is an auxiliary information source.
- LSST-CF model is important for fusing context features from different sources.

Acknowledgements

Joint work supported by Projects on Faculty/Student Exchange and Collaboration Scheme between Higher Education in Hong Kong and the Mainland, Guangzhou Science and Technology Program under grant 1563600439, and Fundamental Research Funds for the Central Universities and also supported by Hong Kong Research Grant Council under General Research Fund. Zhen Li is supported by Lee Shau Kee Postgraduate Fellowship from the University of Hong Kong.
**Goal**

Obtain sharp images from motion blurred stereo videos.

**Challenges**

- **Independent object motion** requires estimating spatially variant, local blur kernels.
- **Inaccurate blur estimation** can lead to severe artifacts in deblurring.
- **Pixels at object boundaries** cause large deblurring errors if they are treated analogously to interior pixels.

**Approach**

- **Stereo video input** enables robust 3D scene flow estimation even in the presence of blur.
- We establish accurate blur kernels from describing the motion with local homographies.
- **Exclusion of object boundaries** from deblurring yields high quality deblurred images for an arbitrary number of moving objects.

**Method**

**Blur Kernel Construction**

- Image formation via local homographies:
  
  \[
  B(x) = \int_{t_0 - \frac{H^t}{2}}^{t_0 + \frac{H^t}{2}} I_{t_0}(t) \exp(-t\xi) \Pr_{\tau}^{-1} dt
  \]

  \[H^t = Pr_{\tau} \exp(-t\xi) Pr_{\tau}^{-1}\]

  Kernels from homographies

**Piecewise Rigid Deblurring**

- **Oversegmentation into planar patches:**

  Non-planar objects with oversegmentation
  Non-rigid motion with oversegmentation

**Object Boundaries**

- Initialize boundary weights with occlusions from scene flow
- Update the weights iteratively:
  \[
  w_{\alpha}(x) = \exp\left(-k_\alpha \|B(x) - A_x I_{t_0}^{n-1}\|^2\right)
  \]

  Our deblurring with boundary weights

**Energy Minimization**

- Use homography-based image formation model
- Add prior on image gradient
- Solve with iteratively reweighted least squares (IRLS) with alternating re-weighting of data term and smoothness term

\[
E(I_{t_0}) = \sum_{x \in \Omega} \left\|w_{\alpha}(x) (B(x) - A_x I_{t_0})\right\|^2 + \alpha \rho(\nabla I_{t_0}(x))
\]

**Experiments**

**Accuracy on Planar Surfaces**

- Even 2D projections of 3D scene flow improve on optical flow-based results
- At times, deblurring with estimated 3D homographies outperforms deblurring with 2D ground truth displacements.

**Benefits of Boundary Weights**

- Iterative down-weighting suppresses ringing and results in accurate deblurring

**Full Stereo Video Deblurring**

- Full framework consistently improves deblurring compared to competing approaches

<table>
<thead>
<tr>
<th>PSNR</th>
<th>Initial optical flow</th>
<th>2D projection of scene flow</th>
<th>3D homographies</th>
<th>Ours (full)</th>
<th>Kim and Lee [2]</th>
</tr>
</thead>
<tbody>
<tr>
<td>apples</td>
<td>20.89</td>
<td>22.43</td>
<td>25.00</td>
<td>29.48</td>
<td>26.14</td>
</tr>
<tr>
<td>bunny</td>
<td>20.72</td>
<td>22.36</td>
<td>22.95</td>
<td>23.20</td>
<td>21.93</td>
</tr>
<tr>
<td>chair</td>
<td>19.03</td>
<td>21.57</td>
<td>22.19</td>
<td>23.36</td>
<td>21.78</td>
</tr>
<tr>
<td>squares</td>
<td>19.60</td>
<td>21.90</td>
<td>22.56</td>
<td>24.58</td>
<td>22.99</td>
</tr>
<tr>
<td>tripod</td>
<td>24.73</td>
<td>24.55</td>
<td>25.22</td>
<td>26.59</td>
<td>23.30</td>
</tr>
<tr>
<td>bottles</td>
<td>29.51</td>
<td>27.58</td>
<td>30.80</td>
<td>31.07</td>
<td>28.33</td>
</tr>
<tr>
<td>office</td>
<td>31.37</td>
<td>30.87</td>
<td>32.45</td>
<td>32.56</td>
<td>29.01</td>
</tr>
<tr>
<td>planar</td>
<td>30.71</td>
<td>31.21</td>
<td>32.33</td>
<td>32.82</td>
<td>30.01</td>
</tr>
<tr>
<td>toys</td>
<td>33.51</td>
<td>33.64</td>
<td>34.89</td>
<td>34.90</td>
<td>31.06</td>
</tr>
</tbody>
</table>

**Code & Further Information**

http://cvilab-dresden.de/research/image-matching/stereovideo/deblurring
anita.sellent@visinf.tu-darmstadt.de
Robust Image and Video Dehazing with Visual Artifact Suppression via Gradient Residual Minimization

Chen Chen and Minh N. Do          Jue Wang
University of Illinois at Urbana-Champaign          Adobe Research

MATLAB Code and more results:
http://web.engr.illinois.edu/~cchen156/dehaze.html

Problem:
Image dehazing aims to remove the haze on the image and recover the original scene. Most existing image dehazing methods tend to boost local image contrast as well as the image artifacts. In this work, we propose a new method to jointly recover the haze-free image while explicitly minimizing possible visual artifacts.

Image Dehazing:

General steps for dehazing:

\[ J(x) = (I(x) - A) / t(x) + A \]

Considering the existence of artifacts in the input image, we should not directly apply the linear haze model above.

Transmission Refinement:

For the transmission refinement, we propose to use the Total Generalized Variation (GTV) with a guided image encoded in the anisotropic diffusion tensor \( D \):

\[
\min \{ \alpha_1 \int |D^{1/2} (\nabla t - w) dx + \alpha_0 \int |\nabla w| dx + \int |t - \bar{t}| dx \}
\]

\( \alpha_1 \) \hspace{1cm} \( \alpha_0 \)

\( \nabla \theta \)

Data fidelity term

GTV regularization

Results:

Fig. 3. Dehazing results of different methods. (a) Input image. (b) Meng et al.’s result [19]. (c) Li et al.’s result [16]. (d) Galdran et al.’s result [9]. (e) Photoshop 2015 dehazing result. (f) Our result.

Fig. 4. Zoomed-in region of Fig. 4. (a) Input image. (b) Meng et al.’s result [19]. (c) Li et al.’s result [16]. (d) Galdran et al’s result [9]. (e) Photoshop 2015 result. (f) Our result without the proposed GRM. (g) Our result. (h) Our E 10.

Gradient residual Minimization:

We observe that the boosted artifacts can be hardly found in the input image. We propose to minimize gradient residual between the input and output image:

\[
\min \frac{1}{2} \int |J(x) - (I(x) - A) / t(x) + A|^2 dx + \lambda \int |E|_0 dx + y \int |\nabla J - \nabla I|_0 dx
\]

Data fidelity term

Artifacts

Gradient residual

Fig. 2. The convergence of proposed method.

Fig. 5. A frame of video dehazing results.

MATLAB Code and more results:
http://web.engr.illinois.edu/~cchen156/dehaze.html
Input: M affinity graphs \( G^{(v)} = (X,W^{(v)}) \), where vertices of the graph \( X = \{x_1,x_2,\ldots,x_N\} \) represent images, and \( W^{(v)} \) is the edge weights of the \( v \)-st graph.

Output: A more faithful similarity measure \( D \).

**Sparse Contextual Activation (SCA)**

A context-sensitive similarity can be defined
\[
\hat{d}_f(x_q, x_p) = 1 - \frac{\sum_{i=1}^N \min \{F_{q,i}, F_{p,i}\}}{\sum_{i=1}^N \max \{F_{q,i}, F_{p,i}\}},
\]
where \( F \in \mathbb{R}^{N \times 1} \) is the sparse contextual activation defined as
\[
F_{q,p} = \begin{cases} 
\exp(-d(x_q, x_p)) & \text{if } x_p \in N_q(x_q) \\
0 & \text{otherwise}
\end{cases}
\]
The key to make it work lies that similar images select same neighbors with same responses.

**Smooth Neighborhood**

Select neighbors smoothly along the underlying manifold
\[
\min \sum_{i<j}^N w_{ij} \|Y_i - Y_j\|^2 + \mu \sum_{i=1}^N \|Y_i - I_i\|^2,
\]
where \( Y_i = [y_{i1}, y_{i2}, \ldots, y_{iN}] \in \mathbb{R}^{1 \times N} \) is the indicator functions of \( x_i \) that describes the probability distribution of its neighbors.

When multiple affinity graphs are available, we want to learn a share indicator
\[
\min_{\alpha} \sum_{i=1}^M \sum_{j=1}^N w_{ij} \|Y_i - Y_j\|^2 + \mu \sum_{i=1}^N \|Y_i - I_i\|^2,
\]
s.t. \( \sum_{i=1}^M \alpha^{(i)} = 1, 0 \leq \alpha^{(i)} \leq 1, \)
where \( \alpha = \{\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(M)}\} \) is the weight of affinity graphs, and \( \gamma > 1 \) is the weight controller.

**Experiments**

- **MPEG-7 shape dataset**
  - Qualitative Evaluation: using Jaccard similarity between two neighborhood sets \( S(x_q, y_p) = \{N_q(x_q) \cap N_p(y_p)\} \)
  - Improving Context-Sensitive Similarity: combining neighborhood selection techniques with SCA to provide retrieval performances

- **Ukbench Image dataset**
  - Improving Context-Sensitive Similarity: combining neighborhood selection techniques with SCA to provide retrieval performances

**References**


**Acknowledgement**

This work was supported in part by NSFC 61573160, NSFC 61429201 and China Scholarship Council. This work was supported in part to Dr. Qi Tian by ARO grants W911NF-15-1-0200 and Faculty Research Gift Awards by NEC Laboratories of America and Blippar.
Title Generation for User Generated Videos
Kuo-Hao Zeng*, Tseng-Hung Chen*, Juan Carlos Niebles*, Min Sun*
National Tsing Hua University* Stanford University*

Title Generation --- Neural Clickbait [3] ---

Captions: A man riding on bike. A man does a stunt on a bmx bike.
Video title generation aims to produce a title sentence describing the most salient event given a typical 1 minute user-generated video (UGV). While a caption describes a video as a whole and tends to be more generic.

Contributions

Videos

Method 1: Highlight Sensitive Captioning (Sec. 4.1)

Method 2: Sentence Augmentation (Sec. 4.2)

Dummy video observation π

Contribution 1: Highlight Sensitive Captioning
We combine a highlight detector with video captioners [1, 2] to train models that can jointly generate titles and locate highlights.

Contribution 2: Sentence Augmentation
We propose a novel and generally applicable method to train an RNN model with both video-sentence pairs and sentence-only examples, where sentence-only examples are either the description sentences or additional sentences on the web.

Highlight Sensitive Video Captioner
We use a bidirectional single-layer LSTM model with binary classification strategy to discriminate highlight or non-highlight for each video. Higher highlight probability means higher chance to be detected as highlight clip.

Sentence Augmentation
To generate catchy or diverse titles, we propose sentence augmentation method to tackle it. It also deals with the rare word problem in the testing set.

VTV Dataset http://ug-video.com
We collected a large-scale Video Titles in the Wild (VTV) dataset of 18100 automatically crawled user-generated videos and titles.

Quantitative Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Variant</td>
<td>B01</td>
<td>B02</td>
</tr>
<tr>
<td>Vanilla</td>
<td>9.3</td>
<td>3.7</td>
</tr>
<tr>
<td>HL-1</td>
<td>9.2</td>
<td>4.1</td>
</tr>
<tr>
<td>HL</td>
<td>11.4</td>
<td>4.9</td>
</tr>
<tr>
<td>Vanilla+Desc.</td>
<td>7.0</td>
<td>2.5</td>
</tr>
<tr>
<td>Desc. Aug.</td>
<td>10.8</td>
<td>4.6</td>
</tr>
<tr>
<td>Aug. Web</td>
<td>11.0</td>
<td>4.7</td>
</tr>
<tr>
<td>HL+Aug. Web</td>
<td>11.7</td>
<td>5.1</td>
</tr>
</tbody>
</table>

Human Evaluation
We ask 7 subjects to conduct the blind test. All subjects do not know which sentence is predicted by S2VT or our method. Among 1000 videos, 508 (51%) videos are selected as `ours is better`, 405 (40%) videos are selected as `S2VT is better`, and the remaining 87 (9%) videos are selected as `on par`.

[3] https://twitter.com/alexjc/status/769076592396083202
Natural Image Matting using Deep Convolutional Neural Networks

Donghyeon Cho, Yu-Wing Tai, In So Kweon
KAIST 1, SenseTime 2

Introduction

Motivation
- There is a synergistic effect between local and nonlocal matting methods.
- So far, there are no effective ways to combine these two kinds of methods without losing the advantages of both methods.

Objective
- Effectively combine alpha mattes of local and nonlocal methods using deep CNN model to reconstruct higher quality alpha mattes than both of its inputs.
- We choose the closed form matting [1] and KNN matting [2] as representative methods for local and nonlocal methods, respectively.

Experiments

Evaluation

Various inputs

The number of layers

Qualitative results

Real world results

Limitation

Conclusion

- We introduce a deep CNN model for natural image matting.
- Our deep CNN model can effectively combine alpha mattes of local and nonlocal methods to reconstruct higher quality alpha mattes than both of its inputs.
- Our deep CNN method demonstrates outstanding performance.

Double-Opponent Vectorial Total Variation
Freddie Åström and Christoph Schnörr
Heidelberg Collaboratory for Image Processing
Heidelberg University, Germany
freddie.astroem@iwr.uni-heidelberg.de

Introduction
We present a new vectorial total variation method that addresses the problem of color consistent image filtering.
• Our approach is inspired from the double-opponent cell representation.
• Existing methods of vectorial total variation regularizers have insufficient coupling between the color channels and thus may introduce color artifacts.
• We propose a novel coupling between the color channels via a pullback-metric from the opponent space to the data (RGB) space.

Channel mixing
• Modeling psychophysical effects of color in applications is a highly non-trivial problem and many spaces have been proposed, e.g., RGB, sRGB, HSV, YPbPr, YCbCr, CIELAB.
• We consider the double-opponent color space (see Gao et al. [2013]; Land [1983, 1986], Ehner [2007]),
\[ \mathbf{O} = \begin{pmatrix} 1/\sqrt{\pi} & 0 & 0 \\ 0 & 1/\sqrt{\pi} & 0 \\ 0 & 0 & 1/\sqrt{\pi} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix} \]
(1) Denote by the linear mapping \( \mathbf{O} : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \), \( u = (r, g, b) \) \( \rightarrow \mathbf{O}u = \alpha \begin{pmatrix} c_1 & c_2 & c_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{u} \) as the transformation from the RGB color space to the double-opponent space.
(2) The non-linear mapping to the hue (h), saturation (s) and lightness (l) representation of the opponent space is given by \( \mathbf{D} : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \), \( u = (L, a, b) \)
Let \( \mathbf{L} = \begin{pmatrix} l \end{pmatrix} \), \( \mathbf{b} = \begin{pmatrix} b \end{pmatrix} \), \( \mathbf{h} = \begin{pmatrix} h \end{pmatrix} \)
(3) Let \( \varphi: \mathbb{R}^3 \rightarrow \mathbb{R}^3 \) denote the composition of the linear opponent transform and the mapping \( \varphi = (a, b) \rightarrow (L, a, b) \) just discussed above, then we define \( \varphi: u \rightarrow \varphi(u) = \mathbf{O}u \).
(4) The Euclidean inner product \( \langle \cdot, \cdot \rangle \) on the Lαs-space induces via \( \varphi \) the pullback metric on the RGB-space (induced by \( \gamma \)) is given by
\[ \gamma(u) = \langle \mathbf{D}(u), \mathbf{D}(u) \rangle = \begin{pmatrix} \mathbf{a}^T \\ \mathbf{b}^T \end{pmatrix} \mathbf{G}(\mathbf{u}) \mathbf{G}(\mathbf{u})^T \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \]
The Jacobian \( \mathbf{D} \) and the corresponding metric tensor \( \mathbf{G} \), reads
\[ \mathbf{D}(u) = \begin{pmatrix} 1/\sqrt{\pi} & 0 & 0 \\ 0 & 1/\sqrt{\pi} & 0 \\ 0 & 0 & 1/\sqrt{\pi} \end{pmatrix} \begin{pmatrix} \alpha & \beta & \gamma \\ -\beta & \alpha & 0 \\ 0 & 0 & \alpha \end{pmatrix} \]
where \( \alpha = (h \cdot g - r \cdot b - g \cdot r)^2 \), \( \beta = (h \cdot g - r \cdot b - g \cdot r - 2 \cdot r)^2 \), and \( \mathbf{G}(\mathbf{u}) \) has non-normalized eigenvectors \( \lambda_1, \lambda_2, \lambda_3 \) and corresponding eigenvalues \( \Lambda = \frac{1}{2} \mathbf{G}(\mathbf{u}) \mathbf{G}(\mathbf{u})^T \)
(5) The channel coupling term is identified from \( f_2^T \), i.e.,
\[ \gamma(u) = f_2^T(h - r)^2 + \frac{1}{\sqrt{\pi}} \left( \beta \frac{1}{2} - r \cdot g + g^2 - gr + r^2 \right) \]

Proposed model
• The Energy. For a gray-scale image \( u : \Omega \rightarrow \mathbb{R} \) defined on a domain \( \Omega \subset \mathbb{R}^2 \), the total variation measure is
\[ TV(u) = \int_{\Omega} \left| \nabla u \right| = \sup \left\{ \int_{\Omega} \left| \nabla u \right| \right\} \]
A function \( u \in L^1(\Omega) \) belongs to the space of functions of bounded variation \( BV(\Omega) \) if
\[ \| u \|_{BV(\Omega)} = \| u \|_{L^1(\Omega)} + TV(u) < \infty \]
For a color image \( u : \Omega \rightarrow \mathbb{R}^3 \) we propose the minimization problem:
\[ \min_{u} \left\{ E(u) = \frac{1}{2} \langle (Ku - g)^2 \rangle^T + \alpha \sum_{n=1}^{3} TV(u_n) + \beta \lambda_{\text{opp}}(u) \right\} \]
the color channel coupling term is
\[ \lambda_{\text{opp}}(u) = \int_{\Omega} \left| \nabla u \right| = \sup_{\varepsilon > 0} \left\{ \int_{\Omega} \left| \nabla \left( \varepsilon \mathbf{D}(u) \right) \right| \right\} \]

Lemma 1 (Invariance and convexity). \( \lambda_{\text{opp}}(u) \) is rotationally and intensity invariant, 1-homegeneous and convex.

Lemma 2 (Bounded variation). Let \( u \in BV(\Omega; \mathbb{R}^3) \) then \( \nabla u \in BV(\Omega; \mathbb{R}^3) \).

Theorem 1 (Uniqueness and existence of solution). Let \( g \in L^2(\Omega; \mathbb{R}^3) \) and \( u \in BV(\Omega; \mathbb{R}^3) \), then there exists a unique minimizer \( u^* \) of \( E(u) \).

• Optimization. We use a Split-Bregman algorithm (Goldstein and Osher [2009]):
\[ \psi^{k+1} = \psi^k + \varphi(D\psi^k) \frac{1}{\lambda} (D\psi^k - g) + \varphi(D\psi^k) \frac{1}{\lambda^2} \left( D\psi^k - g \right)^2 \]
and solve
\[ u^{k+1} = \frac{1}{\lambda} \left( u^k - \varphi(Du^k) \frac{1}{\lambda} (D\psi^k - g)^2 \right) \]
followed by two shrinkage updates for \( \psi^{k+1}, \psi^{k+2} \).

Color structure
• Illustration. Color structure in a natural image and isoluminant discs as extracted by \( \gamma \).

Proposition 1. \( \gamma \) is invariant to intensity shifts and depends quadratically on color change.

Results
• Experimental setup. The opponent color components are corrupted with Gaussian noise.

<table>
<thead>
<tr>
<th>Original</th>
<th>Noisy (5%)</th>
<th>Corrupted (20%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNRT</td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>Original</td>
<td>16.30/21.87</td>
<td>0.83/0.72</td>
</tr>
<tr>
<td>Noisy</td>
<td>28.2/0.75/4.01</td>
<td>26.5/0.67/3.85</td>
</tr>
<tr>
<td>Corrupted</td>
<td>32.8/1.93/6.76</td>
<td>31.9/1.30/±0.72</td>
</tr>
</tbody>
</table>

• Only our OTV approach produces sharp color transitions and a result similar to the original noise free image.

<table>
<thead>
<tr>
<th>Original</th>
<th>Result</th>
<th>Corrupted</th>
<th>Original</th>
<th>Result</th>
<th>Blurred</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM3DS (Babel '07)</td>
<td>26.3/0.61/5.60</td>
<td>32.6/1.93/6.76</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Further results
**Task**

- Goal: Identify both visible and occluded portions of each object instance.

- Solving this task would enable sophisticated occlusion reasoning. Given the modal and amodal masks, we can infer:
  - Amodal bounding box
  - Presence of occlusion
  - Extent, boundary and region of occlusion
  - Relative depth ordering

- Consequently, amodal instance segmentation is harder than these tasks.

**Challenge**

- Lack of supervised training data: no amodal segmentation annotations are available.
- We must train a model without supervised training data.
- Observation: Hard to undo occlusion, but easy to generate occlusion.
- We train the model on synthetic occlusions generated from standard (modal) instance segmentation data.

**Generating Training Data**

- Step 1: Randomly crop a patch from an image and its corresponding modal mask.

- Step 2: Overlay random objects and retain the original mask (which is now amodal).

- Step 3: Rescale the composite patch and sample a modal bounding box.

- Generated samples:

**Method**

- The neural net architecture is the same as that used by Iterative Instance Segmentation [1], which is based on the hypercolumn architecture.

- Initially, we use a detection system and a modal instance segmentation system to produce modal bounding boxes and heatmaps for each object.

- In each iteration, our system takes the patch inside the current amodal bounding box and the modal heatmap as input and predicts the amodal heatmap for a larger patch.

- Based on the heatmap prediction, it decides how to expand the amodal bounding box in the subsequent iteration.

**Results**

- On Small Set of Annotated Images:
  
<table>
<thead>
<tr>
<th>Method</th>
<th>mAP at 50% IoU</th>
<th>mAP at 70% IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN + IO [1]</td>
<td>34.1</td>
<td>14.0</td>
</tr>
<tr>
<td>Faster R-CNN + AIS</td>
<td>45.2</td>
<td>22.6</td>
</tr>
</tbody>
</table>

- On PASCAL 3D+ (Rigid Objects):

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP at 50% IoU</th>
<th>mAP at 70% IoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN + IO [1]</td>
<td>37.4</td>
<td>15.9</td>
</tr>
<tr>
<td>Faster R-CNN + AIS</td>
<td>44.0</td>
<td>20.9</td>
</tr>
</tbody>
</table>

**References**

Perceptual Losses for Real-Time Style Transfer and Super-Resolution
Justin Johnson, Alexandre Alahi, and Li Fei-Fei
Stanford University

Overview
- Image transformation problems typically train feedforward networks with a per-pixel loss between the output and ground-truth
- Recent work [1,9] generates images by optimizing perceptual loss functions defined using features of pretrained CNNs
- Combining these ideas, we use perceptual losses to train feedforward networks
- For style transfer, we train feedforward networks that give similar results as Gatys et al [1] but are up to 1000x faster
- For single-image super-resolution, replacing a per-pixel loss with a perceptual loss gives visually pleasing results

Model
During training we optimize the parameters of the image transform net. The loss network is used to compute content and style losses; it has been pretrained for image classification and remains fixed as the transform net is trained.

Perceptual Content Loss
The content loss between two images is the Euclidean distance between feature maps in the loss network. For style transfer we minimize the content loss between the network output and the content target. For super-resolution we minimize the content loss between the network output and the ground-truth image.

Perceptual Style Loss
The Gram matrix is the uncentered covariance of features from the loss network; the style loss between two images is the Euclidean distance between Gram matrices. Style loss is not used for super-resolution.

Code on GitHub
- Code release includes:
  - Pretrained models
  - Additional models with Instance Normalization [8]
  - Running on new images
  - Training new models
  - Real-time webcam demo

Perceptual Style Loss
![Style reconstruction of an image by minimizing style loss at various layers](image)

Style Transfer Results
- The Starry Night
  - Vincent van Gogh
  - 1889
- The Great Wave
  - off Kanagawa
  - Katsushika Hokusai
  - 1830 - 1832
- The Muse
  - Pablo Picasso
  - 1935
- Composition VII
  - Wassily Kandinsky
  - 1913

Gatys et al [1] Ours

x4 Super-Resolution Results

Gatys et al [1] Ours (pix) Ours (feat)

x8 Super-Resolution Results

Gatys et al [1] Ours (pix) Ours (feat)

User Study: For each image in BSD100, five workers on AMT were asked results from two methods and picked the one they preferred.

References
[7] Shkel and Brook, 2006
**Polysemous codes**

Overview & Take-home message

- **Polysemous codes**
  - offer the best of two popular classes of indexing methods
  - design inspired by channel-optimized vector quantization literature
  - state-of-the-art by a significant margin on billion-sized benchmarks

Outcome:
- best reported results on SIFT1B [5] and Deep1B [4]
- application to building a kNN graph of 95 millions images → in a few machine-hours

Overview & Take-home message

- **Polysemous codes**
  - offer the best of two popular classes of indexing methods
  - design inspired by channel-optimized vector quantization literature
  - state-of-the-art by a significant margin on billion-sized benchmarks

Outcome:
- best reported results on SIFT1B [5] and Deep1B [4]
- application to building a kNN graph of 95 millions images → in a few machine-hours

Approximate nearest neighbor search

$\mathbf{x} \in \mathbb{R}^d$

**Indexing**

$y_1, y_2, \ldots, y_n \in \mathbb{R}^d$

**Result:** $k = \arg \min_{i \in [1..n]} \| x - y_i \|^2$

Efficient indexing in these vector spaces enables:
- searching similar content in billion-sized image collections
- clustering
- proximity graph → the pictures around the poster are paths in the YF100M similarity graph

Motivation: Taking the best of binary and quantization-based approaches

Binary and quantization codes are often presented as competitors. → yet they both have advantages and drawbacks:

- **Comparison (1:1)**
  - PO
  - 1190M comparisons / s
  - precision=0.143
  - Multi-index hashing
  - high memory overhead

- **Binary**
  - need quantizer centroids
  - 222M comparisons / s
  - precision=0.442
  - With an inverted file
  - low memory overhead

Polysemous codes offer both interpretations:
- can be used as binary codes for speed
- or as quantized codes to get a better distance estimation

Channel Optimized Vector Quantization [2]

Shannon’s separation theorem states that optimality of transmission
- can be obtained by separating compression from error correction
- yet under assumptions (error metric, exclude time-varying channels, etc)

Back to the 90’s:
- renewed interest in Joint source-channel coding
- compression strategies assuming a residual error rate

Channel-optimized vector quantizers, e.g., “Pseudo-Gray Coding” [2]
- minimize the overall expected distortion (both from source and channel)
- typically, optimize the index assignment:
  - permuting centroids indexes gives the same quantization accuracy

Polynomial codes: Design

**Goal:** optimize the assignment of centroids to binary codes

**Objective function:** (see also ranking loss in the paper)

the Hamming distance between permuted indices matches

$$\pi^* = \arg \min_{\pi} \sum_{i \neq j} \left[ h(\pi(i), \pi(j)) - f(d(c_i, c_j)) \right]^2$$

optimize permutation over all 256^2 pairs of centroids

- weighting to favor nearby centroids

Pre-filtering based on Hamming distance:

$$q_c = \text{encode}(x)$$

$\text{encode} \{ \}$

for $i = 1..n$

$bc = \text{ycodes}[i]$

get database vector’s code

if $\text{hamming}(qc, bc) < \text{threshold}$

$\text{nearest_neighbor, } d_{\text{min}} = i, d$

true distance

most comparisons are filtered on Hamming distance (XOR+popcnt): 2-stage comparison

Experiments: speed/accuracy/memory usage tradeoffs

Effect of polysemous filtering:

(SIFT1M dataset, $M = 16$)

Comparison to the state-of-the-art on SIFT1B [5] and Deep1B [3]

- Deep1B (8 bytes per code)
  - hardware: Titan X
  - time (ms)
  - Polysemous 1 thread: 0.349
  - 20 threads: 0.035

- Deep1B (8 bytes per code)
  - hardware: Titan X
  - time (ms)
  - Polysemous 1 thread: 0.349
  - 20 threads: 0.035

kNN graph of a collection

Brute-force kNN graph on YF100M images:
- takes 8 hours on 20 cores (extracting image descriptors: 3 days on 4 GPUs)
- 80 GB RAM for 100 neighbors per image

Simple algorithms to:
- find modes of the graph (typical images) by applying power iterations.
- find paths between pairs of images using Dijkstra’s algorithm (minimize the maximum edge distance on the path).

References

Jensen Bregman LogDet Divergence Optimal Filtering in the Manifold of Positive Definite Matrices

Yin Wang, Octavia Camps, Mario Sznaier, Biel Roig Solvas

Motivation

General Linear System Model Estimation

JBLD

Observations

Covariance Feature

Feature Update Model?

Tracking under occlusion

Experiments

YouTube Video Experiment

Mean Estimation Error

Running Time (sec per 100 frames)

This work was supported in part by NSF grants IIS 1358145 and ECCS 1561103, AFOSR grant FA9550-15-1-0302; and the Air Force DOD Center of Excellence under Award Number 2013-ST-061-E0001.
INTRODUCTION

Abstract:
• This paper presents a new reflection symmetry detection method extracting robust Appearance of Structure (AoS) features.
• AoS is located on edges of optimal scale with maximum curvature value and describes the appearance of edges.
• The description of AoS is constructed as 4-dimensional histogram encoding relative location and shape of edges inside local regional neighborhood.
• Comparative quantitative and qualitative evaluation is provided on two public symmetry datasets.
• Human perception based evaluation is suggested.

Contributions:
• New Appearance of Structure (AoS) feature
• Application of AoS feature to Symmetry Detection Task
• Quantitative and Qualitative evaluation on two public datasets
• Human perception based Evaluation

FEATURE DETECTION AND DESCRIPTION

AoS Detection:
For each edge point:
• edge segment is extracted
• curvature is calculated
• orientation of edge point is calculated
Key point is edge point with maximum curvature

AoS Description:
• Key point
  • describes relative location and shape of edges inside local region (neighborhood)
  • collectively, stores votes from edge points in a 4D Histogram

SYMOMETRY DETECTION

Key point mirroring example:
• \( \gamma \) and \( \delta \) have 8 bins (each)
• \( \alpha \) and \( \beta \) have 11 bins (each)

Description mirroring example in two 3-dimensional parameter spaces:
• Symmetry line detection
  • key points are matched to mirrored key point
  • KNN is used in matching

EXPERIMENTAL RESULTS

CVPR2013 dataset results (comparison with Loy et al. [6]):

CVPR2015 Images (Atadjanov and Lee [11]):

Failure Cases:
• Nature images
• Background clutter
• Small Foreground

Human Perception based Evaluation:
• Ground truth is not perfect
• Human perception based evaluation on single symmetry category of CVPR2013 dataset
Faceless Person Recognition; Privacy Implications in Social Media
Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany

Motivation
- How much private information can be exposed from social photos via computer vision?
- How robust are the state of the art person recognisers to head blur?
- Which actions can users take to protect their privacy?

Challenges in Analysis
- Can only lower bound on the performance of the best corporate systems, due to a limited access to the large scale private user databases.
- How to simulate users with varying degrees of privacy sensitivity?
- How to aggregate personal information spread across multiple photos?

Setup for Analysis
- Person recognition in social media.
- Closed world assumption: Recognise from a finite set of identities (200–600).
- GT head boxes are given on all the instances.
- Fuse information from non-tagged instances in the same album and < 10 tagged instances per identity.
- Consider multiple identity protection scenarios.
- Dataset: Person In Photo Albums (PIPA) [1]

Identity Protection Scenarios

<table>
<thead>
<tr>
<th>Number of tagged photos &amp; amount of head obfuscation</th>
<th>Who?</th>
<th>In the same album</th>
<th>Tagged examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1^i = 2$ Many tagged heads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_1^i = 1$ Few tagged heads</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_2$ Obfuscate query head</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_3$ Obfuscate every head</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Head obfuscation types
- Fully visible
- Blur
- Black fill-in
- White fill-in

Domain shift [2]
- Within events: Similar clothing.
- Across events: Changed clothing.

Conclusion in a Nutshell
1. State of the art person recognisers are robust to common identity protection measures.
2. Further performance boost from 1) adapting system to obfuscation patterns and 2) jointly reasoning across photos.
3. Even in the most protective scenario (no identity tag in the same event photos, all heads obfuscated), achieve 12x above naive guess.

Faceless Person Recognition
$$
\arg \max_y \frac{1}{V} \sum_{i \in V} \phi_i(y_i | X_i) + \frac{1}{|E|} \sum_{e \in E} \psi_e(y_e | X_e, X_i)
$$

Unary: single person recogniser. $\phi$:
- Identity probabilities for a single person.
- State of the art CNN full-body recogniser [2].
- Fine-tuned for obfuscation patterns.

Pairwise: person pair matcher. $\psi$:
- Match probabilities for person pairs.
- Siamese network trained for matching.
- Fine-tuned for obfuscation patterns.

Quantitative Results
Identification accuracy versus tag rate.

<table>
<thead>
<tr>
<th>Number of tagged photos:</th>
<th>1.25 tags / person $\rightarrow$ still far better than naive baseline.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of head obfuscation:</td>
<td>Within events: ineffective way of protection.</td>
</tr>
<tr>
<td>Head obfuscation types:</td>
<td>Black $\Rightarrow$ White $\Rightarrow$ Blur $\Rightarrow$ Visible.</td>
</tr>
<tr>
<td>Domain shift:</td>
<td>Recognition system struggles more across events.</td>
</tr>
<tr>
<td>Take-away:</td>
<td>Make sure no tagged heads exist for the event where you want protection.</td>
</tr>
</tbody>
</table>

Qualitative Results

References
Segmental Spatiotemporal CNNs for Fine-grained Action Segmentation

Colin Lea, Austin Reiter, Rene Vidal, Gregory D. Hager

Motivation

Fine-grained activity recognition will propel next generation assistive technologies

Goal: Predict a sequence of actions within video

1) Improved Dense Trajectories (IDT) + bag of Words do not work well on FGAR
2) Spatial & Spatiotemporal CNNs have been proposed, but IDT or CNN+IDT is typically superior [Sun ICCV15, Heilbron CVPR15, Simonyan ICLR15, Jain CVPR15, Tan IC CV15, Karpathy CVPR14, …]
3) Current action localization results are poor, implying that these models capture scene information but not the essence of what defines an action

Context

Situated environments: We assume the camera is static so we can capture geometric relationships among objects

Fine-grained actions: Differentiate between similar actions such as cutting, mixing, and pouring.

Approach

Factorize video into spatial & temporal components

- A spatial network captures objects, locations, and relationships
- A temporal network captures relationships change with time

Predict action segments (not frames)

- Introduce constrained segmental inference algorithm
- Ensure temporal consistency & prevent over-segmentation

Results & Visualization

Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World @ MICCAI M2CAI workshop

Additional applications of this work

- Temporal Convolutional Networks
- Generalizes our temporal convolutions
- Unifies our unary and temporal models
- New results on GT EA, 50 Salads & JIGSAWS

See our follow up ECCV workshop paper

Temporal Convolutional Networks @Brave New Ideas for Motion Representations

- Generalizes our temporal convolutions
- Unifies our unary and temporal models
- New results on GT EA, 50 Salads & JIGSAWS

### Constrained Segmental Inference

**Issue:** RNNs, CRFs, HMMs, etc. predict actions per-frame

**Solution:** Jointly compute start/end/class per-segment \( P_j = (t_j, e_j, c_j) \)

Maximize score \( E(S, P) \) such that there are no more than \( K \) segments

**Objective**

- \( E(S, P) = \sum_{m=1}^{M} f(X, s_m, e_m, c_m) + p(c_{m-1} - c_m) \)

**Recursion**

- **Algorithm 1:** Our Semi-Markov Forward Pass

**Complexity**

- \( K \)-segments (ours): \( O(KTC^2) \)
- Segmental Viterbi: \( O(DTC^2) \)

**Speedup** (Typically \( K \ll D \))

- 50 Salads: 35x-1900x
- JIGSAWS: 50x speedup

*There is a technical error in the ECCV paper. See arXiv for updated equations.*

### Spatial component (VGG-like)

- Apply sequence of spatial units (3x3 conv, ReLU, max pool)
- One intermediate layer of latent fully connected states

Note: pre-trained CNNs were found insufficient for situated tasks

### Temporal component

- Apply 1D convolutions over activations \( \{h_1, \ldots, h_t\} \) to capture how the scene changes over time


3 spatial units with 32, 64, 128 filters each. 64 temporal units.

Input: RGB + Motion History Image (MHI) for each frame

Output: Probability of each action occurring at each frame

**JIGSAWS**

- Spatial component: VGG-like
- Temporal component: 1D convolutions

### Results & Visualization

**50 Salads**

Spatial Models (Edit Accuracy)

- VGG: 28.8, 33.3
- S-CNN: 28.0, 33.0

- Spatial Models (Edit Accuracy) + Temporal CNN

- VGG: 28.0, 33.0
- S-CNN: 28.0, 33.0

High dimensional spatial activations are hard to interpret, so we threshold our model on low-dimensional accelerometer data on 50 Salads.

### Additional applications of this work

- Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World @ MICCAI M2CAI workshop

- Temporal Convolutional Networks
- Generalizes our temporal convolutions
- Unifies our unary and temporal models
- New results on GT EA, 50 Salads & JIGSAWS

---

**Factorize video into spatial & temporal components**

- A spatial network captures objects, locations, and relationships
- A temporal network captures relationships change with time

**Predict action segments (not frames)**

- Introduce constrained segmental inference algorithm
- Ensure temporal consistency & prevent over-segmentation

**Context**

- Situated environments: We assume the camera is static so we can capture geometric relationships among objects
- Fine-grained actions: Differentiate between similar actions such as cutting, mixing, and pouring.

**Goal**

- Predict a sequence of actions within video

**Motivation**

- Fine-grained activity recognition will propel next generation assistive technologies

**Results & Visualization**

- Surgical Phase Recognition: from Instrumented ORs to Hospitals Around the World @ MICCAI M2CAI workshop

---

**Constrained Segmental Inference**

**Issue:** RNNs, CRFs, HMMs, etc. predict actions per-frame

**Solution:** Jointly compute start/end/class per-segment \( P_j = (t_j, e_j, c_j) \)

Maximize score \( E(S, P) \) such that there are no more than \( K \) segments

Compute scores per-segment (vs. per-frame in Segmental Viterbi)

**Objective**

- \( E(S, P) = \sum_{m=1}^{M} f(X, s_m, e_m, c_m) + p(c_{m-1} - c_m) \)

**Recursion**

- **Algorithm 1:** Our Semi-Markov Forward Pass

**Complexity**

- \( K \)-segments (ours): \( O(KTC^2) \)
- Segmental Viterbi: \( O(DTC^2) \)

**Speedup** (Typically \( K \ll D \))

- 50 Salads: 35x-1900x
- JIGSAWS: 50x speedup

*There is a technical error in the ECCV paper. See arXiv for updated equations.*
**Structure from Motion on a Sphere**

Jonathan Ventura  
Department of Computer Science  
University of Colorado Colorado Springs, USA

---

**INTRODUCTION**

- Visual 3D reconstruction requires adequate **baseline** between images.
- However, users of handheld visual SLAM systems typically **rotate** the camera instead of translating it.
- The resulting small translations make structure-from-motion unstable.
- **Idea:** Assume the camera rotates on the surface of a sphere to constrain the structure-from-motion problem.
- For example, with a handheld camera, the shoulder is the origin point and the arm is the fixed radius.
- **Possible applications:**  
  - Handheld SLAM initialization  
  - Stereo panorama creation  
  - 3D face scan with selfie stick  
  - Handheld object scanning  
  - Spherical camera gantry

**EPIPOLAR GEOMETRY**

Outward-facing camera pose is

\[ P_{\text{out}} = [R \ -z] \]

where

\[ z = [0 \ 0 \ 1]^T. \]

Relative pose between

\[ P_1 = [R_1 \ -z] \]  
\[ P_2 = [R_2 \ -z] \]

is

\[ P = [R_2 R_1^T \ r_3 - z] \]

where \( r_3 \) is the third column of \( R_2 R_1^T \).

Essential matrix is

\[ E = [r_3 - z] R_2 R_1^T. \]

For inward-facing camera, translation is opposite.

Note that camera absolute and relative pose are determined completely by 3-DOF rotations.

---

**CONCEPT**

The camera rotates on the surface of a sphere, with its optical axis normal to the surface. The camera could face inward or outward.

---

**THREE-POINT RELATIVE POSE MINIMAL SOLVER**

Relative pose is determined by three rotational degrees of freedom  
\( \rightarrow \) need at least three point correspondences

Essential matrix has the form

\[ E = \begin{bmatrix} e_1 & e_2 & e_3 \\ -e_2 & e_1 & e_4 \\ e_5 & e_6 & 0 \end{bmatrix}. \]

Each correspondence between \( u_i \) and \( v_i \) gives a constraint of the form

\[ v_i^T E u_i = 0. \]

Minimal solver approach:
1. Stack linear constraints into an \( N \times 6 \) matrix and find \( 6 \times 3 \) nullspace.
2. Apply independent subset of six non-linear constraints from:

\[ EE^T E - \frac{1}{2} \text{trace}(EE^T)E = 0 \]

3. Solve non-linear constraints using action matrix or hidden-variable resultant.  
\( \rightarrow \) Produces up to four real-valued solutions for \( E \).
4. Select best essential matrix using an extra correspondence.
5. Decompose essential matrix and resolve twisted pair ambiguity using knowledge of inward- or outward-facing cameras.

---

**structure from motion**

1. Pairwise spherical relative pose estimation  
2. Global pose estimation on the sphere via rotation averaging [CG13]
3. Inverse depth bundle adjustment [YG14] on the sphere

Camera parameterization: Three rotation parameters  
Point parameterization: Inverse depth in reference frame  
Objective function: Huber cost function on re-projection error

---

**RESULTS**


---

**REFERENCES**

This material is based upon work supported by the National Science Foundation under Grant No. 1464420.

---

**ACKNOWLEDGMENTS**
Motivation

- LBP has emerged as one of the most prominent texture features due to:
  - its overall computational simplicity
  - its monotonic gray-scale invariance
  - its flexibility
  - its ease of implementation
  - its effectiveness

- A number of new LBP variants continue to be proposed.

- New types of texture features have also emerged:
  - features based on multistage convolutional networks
  - features based on Deep Convolutional Neural Networks (DCNN)

- Some limitations with the performance comparison in recent papers:
  - when comparing the proposed method to LBP, only basic single-resolution LBP methods have been normally considered;
  - experimenting with some small texture datasets;
  - experimenting with differing classifiers and testing protocols;
  - often not using the best sets of parameter values and multiple scales for the comparative methods;
  - very important aspects such as computational complexity and effects of poor image quality are often neglected.

- No systematic performance evaluation has been carried out.

- It is highly pertinent to perform a more comprehensive performance evaluation and fair comparison of LBP approaches against novel challengers from the deep learning domain.

Our Work

- We proposed a new extensive benchmark for measuring the robustness of texture descriptors against different classification challenges:
  - changes in rotation
  - changes in scale
  - changes in illumination
  - changes in viewpoint
  - changes in number of classes (including large-scale)
  - different types of image degradation
    - random Gaussian noise
    - Salt and pepper noise
    - image blurring
    - random image corruption
  - computational complexity

- We derived fourteen datasets from the eight most commonly used texture sources.

- We presented an extensive evaluation of the recent most promising LBP variants and some non-LBP descriptors based on deep convolutional networks.

Conclusions

- The best overall performance is obtained for the MRELBP feature.

- Both micro- and macro-structures are important for texture description.

- A combination of multiple complementary texture descriptors turns out to be more powerful than a single descriptor.

- For textures with very large appearance variations, Fisher Vector pooling of deep Convolutional Neural Networks is clearly the best, but at the cost of very high computational complexity.

- The sensitivity to image degradations and computational complexity are among the key problems for most of the methods considered.

- A good large scale texture dataset like ImageNet is demanded for the texture research community.
Objective: to recognize one million celebrities from their face images and identify by linking to the unique entity keys in a knowledge base

- Not enough effort in determining the identity with disambiguation
- Public datasets are much smaller than private ones

### Benchmark Task

- **Knowledge base**
  - Recognition with disambiguation
  - The linked entity key is associated with rich and comprehensive property information
  - Beneficial to many real applications
- **Celebrities**
  - Represents public interest
  - Rich information available to build large-scale dataset
- **One-million**
  - Top 1M according to web occurrence (born after 1846)
  - Largest classification problem in computer vision

### Experimental Results

- **Baseline performance (AlexNet)**
  - Allow the model to perform rejection
- **Performance of different methods in the Grand challenge**
  - Evaluated by coverage (how many out of 1000 being recognized) at 95% precision

<table>
<thead>
<tr>
<th>Team Name</th>
<th>Coverage@P=0.95</th>
<th>Team Name</th>
<th>Coverage@P=0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Set</td>
<td></td>
<td>Hard Set</td>
<td></td>
</tr>
<tr>
<td>DRNLSR</td>
<td>0.734</td>
<td>CIGIT_NLPR</td>
<td>0.534</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.728</td>
<td>DRNLSR</td>
<td>0.486</td>
</tr>
<tr>
<td>ITRC-SARI</td>
<td>0.707</td>
<td>Baseline</td>
<td>0.442</td>
</tr>
<tr>
<td>CIGIT_NLPR</td>
<td>0.684</td>
<td>faceman</td>
<td>0.330</td>
</tr>
<tr>
<td>ms3rz</td>
<td>0.646</td>
<td>ms3rz</td>
<td>0.260</td>
</tr>
</tbody>
</table>

**Dataset**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Available</th>
<th># of People</th>
<th># of faces</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFW</td>
<td>public</td>
<td>5k</td>
<td>13k</td>
</tr>
<tr>
<td>YFD</td>
<td>public</td>
<td>1.5k</td>
<td>3.4 k videos</td>
</tr>
<tr>
<td>CelebFaces</td>
<td>public</td>
<td>10k</td>
<td>202k</td>
</tr>
<tr>
<td>CASIA-WebFace</td>
<td>public</td>
<td>10k</td>
<td>500k</td>
</tr>
<tr>
<td>MS-Celeb-1M</td>
<td>public</td>
<td>100k</td>
<td>About 8,456k</td>
</tr>
<tr>
<td>Facebook</td>
<td>private</td>
<td>4k</td>
<td>4,400k</td>
</tr>
<tr>
<td>Google</td>
<td>private</td>
<td>800k</td>
<td>100-200m</td>
</tr>
</tbody>
</table>
A 4D Light-Field Dataset and CNN Architectures for Material Recognition

Ting-Chun Wang¹, Jun-Yan Zhu¹, Ebi Hiroaki², Manmohan Chandraker², Alexei A. Efros¹, Ravi Ramamoorthi²
¹University of California, Berkeley ²University of California, San Diego

Motivation
• Light-field images should help recognize materials since reflectance can be estimated
• CNNs have recently been very successful in material recognition
• We combine these two and propose a new light-field dataset since no one is currently available

Example images in dataset

System Overview

Network architectures

Input: Light-field (microlens) image
Method 1: Apply angular filter on (microlens) image
Method 2: Decompose a 4D filter into a spatial and an angular filter

Results

Quantitative results:

<table>
<thead>
<tr>
<th>Architecture</th>
<th>2D</th>
<th>2D avg</th>
<th>viewpool</th>
<th>stack</th>
<th>EPI</th>
<th>angular</th>
<th>4D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>70.2±1.0</td>
<td>70.5±0.9</td>
<td>70.0±1.0</td>
<td>72.8±1.1</td>
<td>72.3±1.0</td>
<td><strong>77.0±1.1</strong></td>
<td><strong>77.0±1.1</strong></td>
</tr>
</tbody>
</table>

Full scene segmentation results

Example advantages of using light-fields

2D

<table>
<thead>
<tr>
<th>paper</th>
<th>sky</th>
</tr>
</thead>
</table>

Light-field (micro-lens)

<table>
<thead>
<tr>
<th>paper</th>
<th>sky</th>
</tr>
</thead>
</table>

Original photo | Printed photo | 2D prediction | LF prediction |
Graph-Based Consistent Matching for Structure-from-Motion
Tianwei Shen, Siyu Zhu, Tian Fang, Runze Zhang, Long Quan
Hong Kong University of Science and Technology

Motivations
- Pairwise image matching in SfM is costly and contains errors.
- Insufficient match pairs may result in disconnected structures or incomplete components. The number of retrieved items k for a query image is hard to determine.
- Redundant pairs containing erroneous matches may lead to folded and superimposed structures.
- This paper presents a graph-based image matching method that tackles the issues of completeness, efficiency and consistency in a unified framework.

Preliminaries
1. Vocabulary Tree [3]: To avoid the costly exhaustive match, image retrieval has been extensively employed as a pre-processing step for large-scale SfM.

2. Ambiguous Structures: Identification and removal of erroneous epipolar geometry is a recent research focus for SfM.

3. Loop Consistency [4]: Chained relative motion should be an identity map, $R_{12}R_{23}R_{31} = I$.

- The problem is casted as a Bayesian inference task.
- Strong assumption on variable independence, as is pointed out by [5].

Methods

The proposed method can be decomposed into three steps illustrated in: a) match graph initialization, b) graph expansion by strong triplets and c) community-based graph reinforcement. Each step is detailed as follows:

- Starts from a minimal spanning tree (MST) based on vocabulary tree ranks
  - The purpose is to quickly chain the views
  - A modified Kruskal's algorithm (online version): reject outliers
  - Edge weight parameterized by rank information given by vocabulary tree: $w(r_{ij}) = \sqrt{\text{Rank}(r_{ij}) + \text{Rank}(c_{ij})}$

- Expand the spanning tree with loop consistency guaranteed
  - Verifying all loops is hard to achieve, even verifying all triplets is $O(n^3)$
  - Generate a consistent match graph in a bottom-up fashion
  - A empirical choice: traversing two steps starting from each node

- Find loop closures by community detection [1]
  - Too sparse connection after triplet expansion
  - Longer loops are not verified
  - Community detection: divide a graph into groups with denser connections inside and sparser connections outside
  - The modularity $Q$ measures the difference of the fraction of intra-community connections between a graph and the random graph: $Q = \sum_{i,j} (A_{ij} - d_i d_j) (K_{ij} - K_i K_j)$

Experimental Evaluations

We demonstrate the superior performance of our method in terms of accuracy and efficiency on benchmark and large-scale Internet datasets, as well as highly ambiguous scenes.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Reviews</th>
<th>Voc25</th>
<th>Voc100</th>
<th>Matched</th>
<th>Voc25</th>
<th>Voc100</th>
<th>Speedup</th>
<th>Reprojected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trevi Fountain</td>
<td>Ours</td>
<td>1190</td>
<td>1803</td>
<td>756</td>
<td>987</td>
<td>1803</td>
<td>5.7</td>
<td>1324</td>
</tr>
<tr>
<td>Image1</td>
<td>Ours</td>
<td>1150</td>
<td>1723</td>
<td>702</td>
<td>964</td>
<td>1723</td>
<td>4.8</td>
<td>1258</td>
</tr>
<tr>
<td>Image2</td>
<td>Ours</td>
<td>1120</td>
<td>1683</td>
<td>684</td>
<td>956</td>
<td>1683</td>
<td>5.3</td>
<td>1300</td>
</tr>
<tr>
<td>Image3</td>
<td>Ours</td>
<td>1100</td>
<td>1656</td>
<td>678</td>
<td>950</td>
<td>1656</td>
<td>5.5</td>
<td>1300</td>
</tr>
<tr>
<td>Notre Dame</td>
<td>Ours</td>
<td>1420</td>
<td>2247</td>
<td>1208</td>
<td>1962</td>
<td>2247</td>
<td>2.7</td>
<td>1763</td>
</tr>
<tr>
<td>Colosseum</td>
<td>Ours</td>
<td>1780</td>
<td>2591</td>
<td>1392</td>
<td>1918</td>
<td>2591</td>
<td>2.6</td>
<td>1840</td>
</tr>
</tbody>
</table>

Ambiguity Datasets:

- Temple of Heaven Dataset:

Large-scale Internet Datasets:

References
**Introduction**

- **Motivation**
  - For user-created VR contents, hand-held 360 VR cameras are released, but it cannot provide stereoscopic image (needs depth) which is essential for realistic VR contents.
- **Objective**
  - Generate an all-around depth for 360 VR camera from 1-second small motion video.
- **Contribution**
  - Unified bundle adjustment with frontal and rear camera whose residual computed on the unit sphere domain, instead of image domain.
  - Sphere sweeping method on the basis of the unit sphere

- **Small motion Bundle adjustment for SPC (360 VR Camera)**
  - Unified bundle adjustment for both frontal and rear camera
  - Cost function is designed on the unit sphere domain, instead of image domain.
  - Omnidirectional cameras have two-projection model which increases the complexity of the cost function (hardly converges with a high-order model).
  - The re-projection error is uniformly mapped on the sphere which is not the case in the image domain because of the no-linear resolution induced by fisheye lenses.

- **Sphere sweeping for dense matching**
  - Warp images via the successive virtual spheres
  - Covers all-around 3D points only with positive depth representation
  - Find a label that has the highest color consistency
  - Use both frontal and rear images for matching

- **Illustration on the sphere sweeping**

- **Comparison on dense matching method**

**Quantitative evaluation & Qualitative result**

- **Cost function is designed on the unit sphere domain, instead of image domain.**

- **Omnidirectional camera model**

**Application**

- **Stereo Images images for VR head-mounted display & anaglyph**

**Discussion**

- **The reconstruction is up to scale when:**
  - Only pure translation or only z-axis rotation
  - Zero baseline between frontal and rear camera
  - General rotation matrix can be adapted to bundle adjustment for spherical sensor
  - Only small angle approximated rotation matrix is working for pin-hole camera [1-3].
  - The proposed bundle adjustment may have the potential to be generalized to any type of motion.

**References**

[1] F. Yu and D. Gallup. 3d reconstruction from accidental motion. CVPR 2014

**Sunghoon Im, Hyowon Ha, Francois Rameau, Hae-Gon Jeon, Gyeongmin Choe, In So Kweon**

Korea Advanced Institute of Science and Technology (KAIST)
On Volumetric Shape Reconstruction From Implicit Forms
Li Wang, Franck Hétroy-Wheleer, Edmond Boyer
Univ. Grenoble Alpes & Inria & LJK, Grenoble, France

**Objective**
Volumetric tessellation of a shape implicitly defined, satisfying accuracy and quality criteria.

Input: Implicit Form
Output: Volumetric Shape

**Contributions**
- An approach to generate Centroidal Voronoi Tessellations (CVTs) from implicit forms.
- A qualitative and quantitative comparison of Marching Cubes- (MC-), Delaunay-, and CVT-based approaches.

**Algorithm**
- **Implicit Form**
- **Initialization**
- **Clipping (a)**
- **Clipping (b)**
- **Clipping (c)**
- **Optimization**

**Dataset**
- Input: Point Cloud
- Implicit Surface

**Accuracy**
Criterion: Hausdorff distance between the implicit surface and the tessellation boundary.

Delaunay | MC | Ours
--- | --- | ---

**Cell Regularity**
Criterion: Dimensionless second moment of the cell.

\[ G_2(I) = \frac{\int |x - \bar{x}|^2 \, dx}{3V^2(I)^{2/3}} \]

**Computation Time (s)**

<table>
<thead>
<tr>
<th>Method</th>
<th>Dancer</th>
<th>Gargoyle</th>
<th>Skull</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>0.6</td>
<td>1.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Delaunay</td>
<td>1052</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>25</td>
<td>165</td>
<td>363</td>
</tr>
</tbody>
</table>

**Conclusion**
- CVT significantly more accurate than MC and Delaunay strategies.
- MC yet the fastest approach.
- **Future work**: reduce computation time (discrete and hierarchical strategies), generalize to other distances.
Multi-attributed Graph Matching with Multi-layer Random Walks
Han-Mu Park and Kuk-Jin Yoon
Computer Vision Lab., Gwangju Institute of Science and Technology, Korea

MOTIVATION

Graph matching
- Graph matching is the process of finding the correspondence or mapping between nodes of two graphs which best preserves attributes of both nodes and edges.
- Quadratic Assignment Problem (QAP)

\[ \hat{x} = \arg \max_x x^T W x \quad \text{s.t.} \quad x \in \{0,1\}^{N \times M}, \forall x_{ij} \leq 1, \forall x_{ij} \geq 0 \]

x: binary assignment vector, W: affinity matrix

Problem statement
- A single attribute is not capable of representing complex properties of image contents.
- Integrating multiple attributes can cause oversimplification and overfitting!
- How can we consider multiple attributes while preserving their characteristics?

EXPERIMENTAL RESULTS

Comparing methods
- RRWM: Cho et al., ECCV'10
- FGM: Zhou et al., CVPR'12
- SM: Leordeanu & Hebert, ICCV'05
- MPM: Cho et al., CVPR'14
- IFPP: Leordeanu & Hebert, ICCV'09
- GAGM: Gold & Rangarajan PAMI'96

Synthetic random graph matching
- A pair of synthetic graphs which contain outliers and deformations
- Each graph has randomly assigned attributes with different variances.
- Each test is iterated 100 times

Table: Experimental results

<table>
<thead>
<tr>
<th>Deformation</th>
<th>Random parameter</th>
<th>Fixed parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>[v = 0 \pm 0.5]</td>
<td>[v = 5, \theta = 10, \gamma = 2]</td>
</tr>
<tr>
<td>Outlier</td>
<td>[v = 9 - 10]</td>
<td>[v = 5, \theta = 10, r = 0.1]</td>
</tr>
<tr>
<td>Attributes</td>
<td>[v = 1 - 22]</td>
<td>[v = 10, \theta = 5, r = 0.2]</td>
</tr>
</tbody>
</table>

Real image matching
- WILLLOW dataset (5 categories: face, motorbike, car, duck, winebottle)
- Appearance attributes: SIFT descriptor difference, color histogram difference
- Geometric attributes: relative distance histogram, relative angular histogram

Table: Experimental results

<table>
<thead>
<tr>
<th>Method</th>
<th>Second level</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRWM</td>
<td>4.22</td>
</tr>
<tr>
<td>MPM</td>
<td>4.20</td>
</tr>
<tr>
<td>FGM</td>
<td>3.97</td>
</tr>
<tr>
<td>SM</td>
<td>3.97</td>
</tr>
<tr>
<td>GAGM</td>
<td>3.97</td>
</tr>
<tr>
<td>IFPP</td>
<td>3.97</td>
</tr>
</tbody>
</table>

Multi-layer structure for multiple attributes
- Adopting the multi-layer structure to preserve each attribute information.
- Multiple attributes are embedded in multiple layers respectively, and the layers are closely linked to each other.

Advantages
- Characteristics of each attribute can be preserved.
- Relationships among attributes can be defined adaptively.

PROPOSED METHOD: Multi-layer random walk matching

Key concept
- The multi-attributed graph matching problem can be solved by computing the random walk centrality of the proposed multi-layer association graph.

Multi-layer random walk graph matching (MLRWM)

- Affinity-preserving random walking (APRW): Generate next distribution according to transition probability (described in the supra-transition matrix P).
- Reweighting for each layer: Apply inflation and bistochastic normalization process to encourage the one-to-one matching constraints for each layer.
- Layer importance computation: Estimate reliability of each layer by computing amount of change during the reweighting process.
- Integration & Diffusion: Remove contradiction among the multiple reweighted distributions.
- Reweighting jump: Merge reweighted distribution and distribution of APRW.
Overview

We present a method that matches local patches from an RGB-D scene against synthetic local object patches to allow for multi-instance and multi-object detection. To facilitate matching, we regress their descriptors using a CNN. We produce state-of-the-art results while being scalable and operating at around 2 Hz.

Local patches & votes

From synthetic views we extract scale-invariant patches and store each together with a 6D vote and its CNN-feature in a codebook.

\[ \text{patch size} = \frac{m}{z} \cdot f \]

Vote filtering

Similar to [1] we conduct a three-stage filtering where we run mean-shift first on the image plane, then in 3D translational space and finally in quaternion space to remove most spurious votes.

Network architecture & training

Here we employ a convolutional autoencoder (CAE) with which we minimize a reconstruction error \[ ||x - y|| \]. It has been trained on many real scene patches and allows to properly reconstruct input from unseen data.

Results

We show our detections on frames of [3] and present scalability.
A Neural Approach to Blind Motion Deblurring

Ayan Chakrabarti (TTI-Chicago)

www.ttic.edu/chakrabarti/ndeblur
github.com/ayanc/ndeblur

"Camera Shake"

Proposed Architecture

**Goal:** Given blurry patch as input, estimate sharp version.

**Our approach:** Train network to output the complex Fourier coefficients of restoration filter.

**Interpretation from Wiener Deconvolution**

\[ Y(\omega) = X(\omega)K(\omega) + \varepsilon \quad \longrightarrow \quad \hat{X}(\omega) = W(\omega)Y(\omega) \quad W(\omega) = \frac{K^*(\omega)S(\omega)}{||K(\omega)||^2 + \sigma^2} \]

To produce filters, network needs to reason both about unknown blur kernel and image PSD.

Use multi-resolution sub-band decomposition to encode input patch.

Limit connectivity of initial layers by locality in frequency. (Freq-domain analogue of conv)

Experiments

Evaluation on [Sun et al., 2013] dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>95%-ile</th>
<th>Max</th>
<th>Success Rate</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>3.01</td>
<td>5.76</td>
<td>11.04</td>
<td>92%</td>
<td>65s (GPU)</td>
</tr>
<tr>
<td>Michaeli &amp; Irani (2014)</td>
<td>2.57</td>
<td>4.49</td>
<td>9.31</td>
<td>96%</td>
<td>91min (CPU)</td>
</tr>
<tr>
<td>Sun et al. (2013)</td>
<td>2.38</td>
<td>5.98</td>
<td>23.07</td>
<td>93%</td>
<td>38min (CPU)</td>
</tr>
<tr>
<td>Xu &amp; Jia (2010)</td>
<td>3.63</td>
<td>9.97</td>
<td>65.33</td>
<td>86%</td>
<td>25% (CPU)</td>
</tr>
<tr>
<td>Schuler et al. (2015)</td>
<td>4.53</td>
<td>11.21</td>
<td>20.96</td>
<td>67%</td>
<td>22% (CPU)</td>
</tr>
</tbody>
</table>

Blurry Input  | Neural Average  | Final Output |

Need to look beyond standard feed-forward architectures.

Joint Face Representation Adaptation and Clustering in Videos

Zhanpeng Zhang1, Ping Luo2, Chen Change Loy1,2, Xiaoou Tang1,2
1Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
2Shenzhen Key Lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China

The Problem

Face clustering in videos is challenging: 1) the various appearance changes as the story progresses 2) various cinematic styles

It is difficult to learn an universal face representation for all videos. Instead, we formulate a joint face representation adaptation and clustering framework.

Problem Definition: 
Input: A set of detected face tracks
Output: Each face is assigned with a label y ∈ {1...K}

- Pre-training as DeepID2+ [1]
- Fine-tuning with pairwise constraints (contrastive loss)

Joint Face Representation Adaptation and Clustering

(a) Feature extraction with a DCN

\[ x = \{ x_f \} \]

input face

representation \( x_f \)

network filters \( W \)

Forward

\[ \text{Model character label } y \text{ in MRF and infer } Y \text{ by maximizing } P(X,Y) \]

back propagation

Pairwise constraints \( C \)

(c) Discovering additional constraints using current features \( X \) and cluster label \( Y \)

Some Priors May help?

- Unary term: Gaussian
- Pairwise term: Face similarity x-spatial and temporal constraints
- Optimization: simulated field algorithm [8]

Diverse clusters contain large noise; clusters with high purity are compact; Faces from the same character are likely to be close; Confidence \( Q \) for a face pair being negative:

\[ Q(\chi) = \frac{1}{1 + \exp(-\chi)} \]

Experiment on Accio-1 (Harry Potter) [2] Movie Dataset

Visualization of different characters’ face representation. (a)-(c): projecting different representations to a 2D space by PCA: (a) raw pixel value, (b) DeepID2+, and (c) our adapted representation. (d): projecting the adapted representation to a 2D space by t-SNE.

B-cubed precision \( P \), recall \( R \), and F1-score \( F \)

Methods & clusters:20 & clusters:12 & clusters:10 & clusters:5 & clusters:20

R-20 F-20 P-20 R-10 F-10 P-10 R-5 F-5 P-5 R-20 F-20 P-20 R-10 F-10 P-10 R-5 F-5 P-5

K-means 240 114 156 202 105 158 280 308 136 401 321 109 355 809 169 612 674 132

K-means-DeepID2+ 543 201 293 264 181 273 581 135 204 394 899 169 612 674 132

DIFFRAC-13 207 109 160 216 209 123 228 208 111 330 395 999 317 422 459

DIFFRAC-DeepID2 507 213 301 226 381 277 607 160 253 622 120 203 612 668 122


BMIF 393 119 174 395 101 151 300 100 113 342 307 112 458 641 574

BMIF-Fiber 583 234 354 251 283 281 604 176 273 567 127 213 712 866 154

BMIF-DeepID2 509 230 312 616 211 314 621 274 372 644 129 214 609 675 153

DeepID2-C2 405 253 365 676 238 352 684 192 300 715 115 255 736 522 226

DeepID2-C2-Beta 657 211 423 685 286 401 698 229 335 723 236 711 154 263

Full model 711 352 471 739 312 439 768 241 308 779 203 322 841 172 284

Experiment on BF0502 [6] and Notting Hill [7]

Example generated clusters on BF0502

Accuracy on BF0502 [5] (left) and Notting Hill [6] (right)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-means</td>
<td>39.31 ± 4.51</td>
</tr>
<tr>
<td>ULDML</td>
<td>41.62 ± 0.00</td>
</tr>
<tr>
<td>PPC</td>
<td>78.89 ± 5.15</td>
</tr>
<tr>
<td>HMMF [8]</td>
<td>59.30 ± 2.73</td>
</tr>
<tr>
<td>WBSSLR</td>
<td>62.76 ± 1.10</td>
</tr>
<tr>
<td>Our method</td>
<td>92.15 ± 0.90</td>
</tr>
</tbody>
</table>

Application to Face Verification

We calculate the L2 distance of the representation to measure the pairsimilarity

(a) ROC of face verification on Accio-1 [2] dataset. The number in the legend indicates the verification accuracy. (b) and (c): negative and positive pairs failed to be matched by DeepID2+ [1] but successfully matched after adaptation by our approach.

Overview
In this paper we study symmetries in polynomial equation systems and how they can be integrated into the action matrix method.
- We generalize the partial $p$-fold symmetry from [1, 2]
- We show how to use multiple independent symmetries.
- We provide a simple method for finding hidden symmetries.
- We show two examples where symmetry allows for more compact solvers.

Prior Work
Consider
\[ \begin{cases} x^2 + y - 2 = 0, \\ x^2y^2 - 1 = 0 \end{cases} \]
\( V = \{(1, 1, 1), (x, y, \varphi^3) : (x, y, \varphi^3) \in V \} \).

Two-fold sign symmetry in x-variable; i.e. \((x, y, \varphi) \in V \Rightarrow (-x, y, \varphi) \in V \).
This type of symmetry was studied in [1, 2].

Definition 1. The polynomial \( f(x, y) \) has a partial \( p \)-fold symmetry in \( x \) if the sum of the exponents for \( x \) of each monomial has the same remainder \( q \) modulo \( p \), i.e.
\[ f(x, y) = \sum_{k \in \mathbb{N}} a_k x^k y^k \Rightarrow q \equiv \Gamma^k a_k \text{ mod } p \forall k. \]

Weighted Symmetries
Consider
\[ \begin{cases} x^2 + y - 2 = 0, \\ xy - 1 = 0 \end{cases} \]
\( V = \{(1, 1, 1), (x, y, \varphi^3) : (x, y, \varphi^3) \in V \} \).

No partial \( p \)-fold symmetry, but \((x, y, \varphi) \in V \Rightarrow (x^2, x^2 \varphi, \varphi^2) \in V \).

Definition 2. The polynomial \( f(x, y) \) has a weighted \( p \)-fold symmetry with weights \( c \in \mathbb{Z}^n \) if the \( c \)-weighted sum of the exponents for \( x \) of each monomial has the same remainder \( q \) modulo \( p \), i.e.
\[ \begin{cases} x^2 + y - 2 = 0, \\ xy - 1 = 0 \end{cases} \]
\[ f(x) = \sum_{k \in \mathbb{N}} a_k x^k \Rightarrow q \equiv \Gamma^k a_k \text{ mod } p \forall k. \]

- Weighted sum of exponents should have same remainder.
- For the example in (3) we have \( c = (1, 2) \), \( p = 3 \).
- Then the solution set \( V \) satisfies
\[ (x_1, x_2, \ldots, x_n) \in V \Rightarrow (x^2, x^2 \varphi, \varphi^2, \ldots, x^2 \varphi^7, \varphi^2) \in V. \]

- Partial \( p \)-fold symmetry corresponds to binary weights.

Block Diagonal Action Matrices
Action matrix will be block diagonal if the action polynomial is invariant to the symmetry:
\[ \begin{cases} x^2 + y^2 - 2 = 0, \\ xy - x = 0 \end{cases} \]
\( V = \{(1, 1, 1), (x, y, \varphi^3) : (x, y, \varphi^3) \in V \} \).

Basis for quotient space given by
\[ B = \{x, y, xy, y^2, y^3\} \]
Grouping basis elements with weighted exponents w.r.t. \( c_x \) and \( c_y \) modulo 2
\[ B_{a,0} = \{y, y^2\}, B_{a,1} = \{y^3, y^4\}, B_{b,0} = \{x\}, B_{b,1,1} = \{x y\}. \]

\[ \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ -2 \end{bmatrix} \begin{bmatrix} x^2 \\ y^2 \\ y \\ y^3 \\ y^4 \end{bmatrix} = \begin{bmatrix} x^2 \\ y^2 \\ y^3 \\ y^4 \\ 1 \end{bmatrix}. \]

Hidden Weighted Symmetries
For \( \alpha \in (-\sqrt{2}, \sqrt{2}) \) consider the following family of polynomial systems
\[ \begin{cases} x^2 + y^2 = 1, \\ x + y = \alpha \end{cases} \]
\( (x, y) \rightarrow (y, x) \)

No weighted symmetries. Change of variables reveals a 2-fold symmetry in one of the variables.
\[ \begin{cases} x = x + y, \\ y = x - y \end{cases} \]
\[ \begin{cases} x^2 + y^2 = 1, \\ \alpha \end{cases} \]

This can be done in general if the solutions are stable under some linear transform. In the paper we present a simple method for finding these symmetries and the correct change of variables.

Weak Perspective-2\(n\)-Points
Pose estimation in scaled orthographic camera
\[ \min_{\mathbf{R}^2} \| \mathbf{RA} - \mathbf{B} \|_F^2 \text{ s.t. } \mathbf{R} \mathbf{R}^T = \mathbf{I} = \mathbf{I}^T, \mathbf{R} \in \mathbb{R}^{2 \times 3}. \]

Quaternion representation for scaled rotation
\[ \mathbf{R}(q) = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_0q_3) & 2(q_2q_3 - q_1q_0) \\ 2(q_1q_2 - q_0q_3) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_1q_3 + q_0q_2) \\ 2(q_1q_3 + q_0q_2) & 2(q_2q_3 - q_1q_0) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix}. \]

Finding minimum by looking at stationary points
\[ \nabla_q f(q) = 0, \text{ where } f(q) = \| \mathbf{R}(q) - \mathbf{B} \|_F^2. \]

Rotation parameterization is invariant under sign change and
\[ \mathbf{R}(q_1, q_2, q_3, q_4) = \mathbf{R}(q_0, q_{0q_3} - q_2, -q_0, -q_2). \]

Perform change of variables
\[ q_0 = \frac{1}{2} \sqrt{q_1^2 q_2^2}, \end{bmatrix} \]

In these new variables the rotation matrix becomes
\[ \mathbf{R}(\hat{q}) = \begin{bmatrix} \hat{q}_0^2 - \hat{q}_1^2 - \hat{q}_2^2 - \hat{q}_3^2 & -2\hat{q}_1\hat{q}_2 + 2\hat{q}_0\hat{q}_3 & 2\hat{q}_2\hat{q}_3 - 2\hat{q}_1\hat{q}_0 \\ 2\hat{q}_1\hat{q}_2 + 2\hat{q}_0\hat{q}_3 & \hat{q}_0^2 - \hat{q}_1^2 + \hat{q}_2^2 - \hat{q}_3^2 & -2\hat{q}_2\hat{q}_3 + 2\hat{q}_1\hat{q}_0 \\ 2\hat{q}_2\hat{q}_3 - 2\hat{q}_1\hat{q}_0 & -2\hat{q}_2\hat{q}_3 + 2\hat{q}_1\hat{q}_0 & \hat{q}_0^2 - \hat{q}_1^2 - \hat{q}_2^2 + \hat{q}_3^2 \end{bmatrix}. \]

Symmetries in original parameterization transformed into two 2-fold partial symmetries
\[ c_j = (1, 1, 0, 0), \quad c_{j'} = (0, 0, 1, 1). \]

References
For color images, machine learning based methods for single image super-resolution are highly successful [6, 11, 16].

- Training data is easy to obtain.

- For depth data, methods are mostly based on energy minimization [5, 7].
- Piece-wise affine surfaces.
- Sharp depth discontinuities.

Code and data is published on GitHub
https://github.com/griegler/primal-dual-networks


Train method end-to-end on huge set of synthetically generated depth maps.

Unroll of 1D primal-dual optimization scheme [3] on top of deep network.
- Joint training with deep network.
- Regularization term parameterized by additional network output.
- Learning of optimization hyper-parameters for each step.

$\min_{w} \frac{1}{K} \sum_{k} L(u^k(f(w, s_k), t_k))$ (HL)

Energy functional with anisotropic TGV as in [7]
- Emphasizes piece-wise affine surfaces.
- Convex optimization problem.

$E(u; f(w, s_k)) = R(u, h(w, s_k)) + \frac{\epsilon_r^2}{2} \| u - g(w, s_k) \|_2^2$
$R(u, h(w, s_k)) = \min \alpha_1 \| T(h(w, s_k)) \|_\infty - v \|_1 + \alpha_2 \| v \|_1$ (2)

- Joint training with deep network.
- Regularization parameterized by additional network output.
- Learning of optimization hyper-parameters for each step.

$\nu_n^{p+1} = \text{proj}(\nu_n^{p} + \alpha_1(T(h(v, s_k))) \nabla u^{p+1} - \nu_n^{p+1}))$
$q_n^{p+1} = \text{proj}(q_n^{p} + \alpha_1(T(h(v, s_k))) \nabla v^{p+1})$
$\nu_n^{p+1} = \nu_n^{p} + \alpha_1(T(h(v, s_k))) \nabla u^{p+1} + \alpha_2 \nabla v^{p+1}$
$v_n^{p+1} = v_n^{p} + \alpha_1(T(h(v, s_k))) \nabla u^{p+1}$
$\nu_n^{p+1} = \nu_n^{p} + \alpha_2 \nabla v^{p+1}$

References and Acknowledgment

Indoor-Outdoor 3D Reconstruction Alignment

Andrea Cohen¹, Johannes Schönberger¹, Pablo Speciale¹, Torsten Sattler¹, Jan-Michael Frahm², Marc Pollefeys¹,³

¹ETH Zürich, Switzerland ²UNC Chapel Hill, USA ³Microsoft

Summary

- Indoor and outdoor reconstructions of buildings get disconnected in SfM.
- Goal: align these models in a plausible way.
- Contributions: new approach to align indoors and outdoors through semantic cues:
  - Detect 3D windows from multi-view images
  - Match 3D windows to generate hypotheses
  - New quality metric to rank the hypotheses.

Overview

- Input: set of disjoint indoor and outdoor SfM reconstructions and the original images
- Axis alignment for each model (using VP).
- 2D window estimation:
  - Rectify original images using x- and y-aligned planes.
  - Apply per-pixel classifier.
  - Facade parsing.
- 3D window detection:
  - Project 2D windows on the x- and y-aligned planes.
  - Detect overlapping windows and compute consensus windows.
- Model alignment:
  - Match indoor and outdoor 3D windows to alignment hypothesis.
  - Evaluate and rank hypotheses.

Alignment

\[ M_{\text{out}} \] set of outdoor models
\[ M_{\text{in}} \] set of indoor models

- Window term
  \[ E_{W}(C_i) < E_{W}(C_j) \]

- Intersection term
  \[ E_{I}(C_i) < E_{I}(C_j) \]

minimize

\[ e = E_{W}(C, W_{3D}) + E_{I}(C, M_{\text{in}}, M_{\text{out}}) \]
subject to

\[ E_{I}(C, M_{\text{in}}, M_{\text{out}}) < \lambda \]

Datasets available at: https://www.cvg.ethz.ch/research/indoor-outdoor/

Results

- Quantitative evaluation: GT generated using hand-labeled windows matched by hand.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Theatre</th>
<th>University</th>
<th>Hall</th>
<th>House-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dimensions (m)</td>
<td>35x26</td>
<td>35x25</td>
<td>40x16</td>
<td>9x9.95</td>
</tr>
<tr>
<td>Error (m)</td>
<td>0.65</td>
<td>0.42</td>
<td>0.19</td>
<td>0.54</td>
</tr>
</tbody>
</table>
The Unreasonable Effectiveness of Noisy Data for Fine-Grained Recognition
Jonathan Krause\textsuperscript{1}, Benjamin Sapp\textsuperscript{2}, Andrew Howard\textsuperscript{2}, Howard Zhou\textsuperscript{2}, Alexander Toshev\textsuperscript{2}, Tom Duerig\textsuperscript{2}, James Philbin\textsuperscript{2}, Li Fei-Fei\textsuperscript{1}
\textsuperscript{1}Stanford University, \textsuperscript{2}Google

Problem
-Fine-grained recognition works well with labels
-But fine-grained labels are expensive
-There are too many fine-grained categories in the world to annotate by hand: 14k birds, 278k butterflies and moths, 941k insects
-How can we scale up fine-grained recognition?

Data
Categories
Birds: 10,982 species
Butterflies: 14,553 species (+moths)
Aircraft: 409 varieties
Dogs: 515 breeds
-Images from Google Image Search

noise
Cross-domain noise: portion of images that are not of any fine-grained category in a given domain. Measure by hand.

Examples of cross-domain noise

Confusion matrices on GT and Web data for CUB categories

Active Learning
Alternative approach for collecting large quantities of fine-grained data.

Experiments
-Use Inception-v3 CNN classifier
-Extensive dedup with ground truth test datasets via \textsuperscript{2}
-YFCC100M data for active learning

Prior work on GT datasets:
CUB: 84.6\% (Xu et al. ICCV'15)
Birdsnap: 66.6\% (Berg et al. CVPR'14)
FGVC: 84.1\% (Liu et al. ICCV'15)
Stanford Dogs: 76.8\% (Sermanet et al. ICLR'2015)

Very Large-Scale Fine-Grained Recognition
-Test on Flickr images w/exact category name matches, deduped with other web images.
-accuracy: Birds (73.1\%), Butterflies (65.9\%), Aircraft (72.7\%)

References

4,224 (+1) categories recognized in this work

Contributions
-Demonstrate feasibility of training models of fine-grained with noisy data from the web and simple, generic, models of recognition.
-Greatly improved recognition performance on four fine-grained datasets without using ground truth training data.
-Scale fine-grained recognition to over 10,000 species of birds and 14,000 species of butterflies and moths.
Overview
We propose a data structure obtained by hierarchically pooling Bag-of-Words descriptors during a sequence of views that achieves average speedups in large-scale loop closure applications ranging from 2 to 20 times on benchmark datasets.

Loop closure is a particular classification task whereby a training set of images is indexed by location, and given a test image one wants to query the database to decide whether the former is present in the latter, and if so return the indexed location.

The challenge with loop closure is scaling. Our goal here is to design a hierarchical data structure that helps speed up loop detection by leveraging on two domain-specific constraints: temporal adjacency, and high precision.

Construction of hierarchy
Given a list of histograms, we build a hierarchy on top of it recursively by pooling adjacent histograms.

Different pooling strategies:
- max-pooling: \( h^p = \max_i (h^i) \) for \( i = 1 \ldots N \)
- sum-pooling: \( h^p = \sum_{i=1}^{N} h^i \)
- mean-pooling: \( h^p = \frac{1}{N} \sum_{i=1}^{N} h^i \)

where \( h^p \) is a parent histogram; \( h^i, k = 1 \ldots K \) are its children. Each histogram has \( N \) bins, which is the vocabulary size.

Performance guarantee
For max-/sum-pooling, no histograms scoring higher than \( \tau \) would be missed according to the following property

\[ I(h, h^p) = \sum_{i=1}^{N} \min(h^i, h^p) \]

Baseline: Linear search accelerated by inverse index structure.

Hierarchical testing
Goal: Given a query histogram and a threshold \( \tau \), find all the histograms scoring higher than \( \tau \) in the database.

Intersection kernel is used to compare histograms

\[ I(h, h^p) = \sum_{i=1}^{N} \min(h^i, h^p) \]

Varying tree topology
We consider both fixed and adaptive tree topology.

In-the-loop test
We use components of ORB-SLAM and build a hierarchy atop its single-layer inverse index based loop detection module.

Extension to image retrieval
Hierarchy leveraging on extra labeling information to speed up image retrieval.

<table>
<thead>
<tr>
<th>structure</th>
<th>pooling time (ms)</th>
<th>speedup mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>inverted index</td>
<td>N/A</td>
<td>1.47</td>
</tr>
<tr>
<td>Random hierarchical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>greedy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>affinity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a) ukbench</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(b) INRIA Holidays</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. A comparison of search in flat and hierarchical structure on ukbench and INRIA Holidays.

Acknowledgements

Fig. 2: (a) Scaling: Timings for concatenated KITTI sequences (approx. 40K images) with 1M and 10K vocabularies. (b) Comparison to ORB-SLAM with and without our data structure. Multiple trials yield nearly identical trajectories with and without our data structure, with no loop closures missed while achieving a 2-3x speedup.

Fig. 3: Timings (top) and precision-recall curves (bottom) of baseline and proposed algorithm with different topologies and pooling strategies on KITTI dataset 00 and 02 using all frames. \( d, d-x \): a hierarchy with \( d \) layers, a branching factor of \( j \) and pooling strategy \( X \). Adaptive sampling: spectral clustering in \( SE(3) \). Regular sampling: sampling at the average rate of adaptive sampling scheme.

Fig. 4. Sample results on the TUM RGB-D dataset using adaptive domain clustering. Adaptive (yellow) improves with more exciting motion.

Varying vocabulary size
Using a larger vocabulary reduces query time, but a sensible speedup is still available with our data structure. A larger vocabulary has finer division of descriptor space compared to a smaller vocabulary but is more sensitive to quantization errors. Vocabulary size should be determined by the task as well as the volume of the data.
1. Introduction

- Depth information is essential in many applications such as autonomous navigation, 3D reconstruction, and human-computer interaction.
- The resolution of depth maps which is provided in a low-cost depth camera is generally very limited.
- We address an upsampling problem in which the corresponding high-resolution (HR) depth map is recovered from a given low-resolution (LR) depth map (and a HR intensity image).

2. Challenges

- Fine structures in the enlarged image are either lost or severely distorted.
- Features in intensity images are often over-transfered to the depth image.

3. MS-Net and MSG-Net

- Only high-freq. part $h(D_I)$ is used for training.
- Upsampled low-freq. part $l(D_h)$ is added back for testing.

### References:


Grid Loss: Detecting Occluded Faces

Michael Opitz  Georg Waltner  Georg Poier  Horst Possegger  Horst Bischof
Graz University of Technology, Institute for Computer Graphics and Vision, Austria

Motivation and Contribution
- Recent benchmark evaluations (e.g., FDDB [2]) show that standard CNN detectors fail on occluded faces.
- State-of-the-art approaches rely on large pre-trained ImageNet models, e.g., [1], and on large datasets with face-attribute annotations to re-rank object proposals [4].
- We show that simple CNN based detectors trained from scratch can outperform these methods, if they handle occlusions properly.
- To this end, we train a part-based CNN detector with our grid loss function.

Supplemental material is available online.

Overview
- 2 convolution layers on top of Aggregate Channel Features.
- Linear pose-specific classifiers on top of the last convolution layer.
- At test time: fully convolutional detection over an image pyramid.
- Regressor to refine location of detected faces.
- To tackle occlusions we look at spatially non-overlapping blocks on the last convolution layer.
- Grid loss optimizes a loss on each of these blocks separately.

Robustness to Occlusions
- Parts are discriminative alone.
- If subset of parts fail (due to occlusion), detection can recover.

<table>
<thead>
<tr>
<th>Method</th>
<th>COFW-HO</th>
<th>COFW-LO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid Loss</td>
<td>0.979</td>
<td>0.998</td>
</tr>
<tr>
<td>Standard Loss</td>
<td>0.909</td>
<td>0.982</td>
</tr>
</tbody>
</table>

Diversity of Learned Features
- More diverse features compared to standard loss.
- Encourages learning discriminative features for all sub-regions.

Generalization Ability
- Better generalization ability, due to larger diversity of features.
- On smaller training set sub-sets the performance gap between Grid Loss and standard loss functions increases.

Benchmark Results
- We achieve state-of-the-art performance on several datasets.

References and Acknowledgments

This work was supported by the Austrian Research Promotion Agency (FFG) project DIANGO (840824).
Real-time, Large-scale Object Detection

**Our goal:** Retrieve objects for a certain category from large image collections immediately and accurately

- search “bicycle” in real-time (100K search in 0.1s)

**Problem:** object detection methods require huge costs, which makes it hard to extend them for large-scale retrieval

**Prior work:**
- BoVW (video google): not effective for category detection
- Aytar+ (CVPR2014): extend DPM using BoVW-like approach

- we achieved significant improvement over both methods

Framework of Large-scale R-CNN

**Offline procedure:**
- object proposals CNN features
  - 2K proposals x 100K images = 200M features
  - compress features by RVQ and store into inverted lists

**Online procedure:**
- query
  - SVM
  - sort by score
  - CAQ improves by 72%: 7/#.:
  - Efficiency over R-CNN: around 5K images, 10M features
  - CAQ(2000), CAQ(1800): 200 categories in ILSVRC detection task

Efficiency over R-CNN: evaluated on PASCAL dataset (around 5K images, 10M features)

<table>
<thead>
<tr>
<th></th>
<th>mAP</th>
<th>time</th>
<th>memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAQ only</td>
<td>52.4%</td>
<td>518.0ms</td>
<td>1636GB</td>
</tr>
<tr>
<td>RVQ only</td>
<td>52.4%</td>
<td>69.5ms</td>
<td>0.64GB</td>
</tr>
<tr>
<td>R-CNN</td>
<td>54.2%</td>
<td>6258ms</td>
<td>1636GB</td>
</tr>
<tr>
<td>Ours</td>
<td>50.7%</td>
<td>24.5ms</td>
<td>0.916GB</td>
</tr>
</tbody>
</table>

250x faster, 106x memory reduction with comparable accuracy

**Our Idea**

**Summary:** extend R-CNN for large scale
- apply SVM to features extracted from all object proposal

**Challenge:** classify millions/billions of features in real-time
- use the techniques of nearest neighbor search

Efficiency: compare methods of inverted index

**CAQ** vs k-means:
- 4 sets of classifiers in CAQ:
  - CAQ(200): 200 in ILSVRC detection task
  - CAQ(1800): 200 in ILSVRC - 20 in PASCAL
  - CAQ(2000), CAQ(1800): CAQ(200), CAQ(1800) x 10 (trained on randomly sampled part of dataset)

CAQ improves by ~10% mAP over k-means

Classier Adaptive Quantization (CAQ)

**Intuition:** k-means is not optimal clustering for our task
- k-means gather close points → positive data stride many clusters

**Key idea:**
- Formulation of CAQ is derived from quantization error:
  - encoding: $i = \arg \min\|Wx - W\mu_i\|^2$
  - codebook learning: minimize
    - similar to k-means with different metric, easy to use
Face Detection with End-to-End Integration of a ConvNet and a 3D Model

Yunzhu Li1,*, Benyuan Sun1,*, Tianfu Wu2 and Yizhou Wang1
1 School of EECS, Peking University
2 Department of ECE and the Visual Narrative Cluster, North Carolina State University
{leo.liyunzhu, sunbenyuan, Yizhou.Wang}@pku.edu.cn, tianfu_zwu@ncsu.edu

Overview

This paper presents a method for face detection in the wild, which integrates a ConvNet and a 3D mean face model in an end-to-end multi-task discriminative learning framework. There are two components:
i) The face proposal component computes face proposals via estimating facial key-points and the 3D transformation parameters for each predicted key-point w.r.t. the 3D mean face model.
ii) The face verification component computes detection results by refining proposals based on configuration pooling.

Figure 1: Illustration of the proposed method (Top), and a sample intermediate and the final detection results (Bottom).

The Proposed Method

Face Representation

A 3D mean face model is represented by a $n \times 3$ matrix, $F^{(3)}$. The 3D transformation parameters $\Theta$ are defined by:

$$\Theta = (\mu, s, A^{(3)})$$

where $\mu$ represents a 2D translation (dx, dy), $s$ a scaling factor, and $A^{(3)}$ a $3 \times 3$ rotation matrix. We can compute the projected 2D key-points by:

$$\pi(F^{(3)}) = \mu + s \cdot \pi(A^{(3)} \cdot F^{(3)}),$$

where $\pi()$ projects a 3D key-point to a 2D one.

ConvNet Architecture

Referring from Figure 1, the ConvNet is consisted by:

- Convolution, ReLu and MaxPooling Layers.
- An Upsampling Layer implemented by deconvolution.
- A Facial Key-point Label Prediction Layer. Samples are shown in Figure 2.
- A Face Proposal Layer. Samples are shown in Figure 3.
- A Key-point based Configuration Pooling Layer.
- A Face Bounding Box Regression Layer.

Figure 2: Sample detection results in the FDDB and the corresponding heat map of facial key-points.

End-to-End Training

During training, the loss are three-folds:

- The Classification Softmax Loss of Key-point Labels,
  $$L_{cls} = - \sum \log(p^\ell),$$  
  where $\ell$ is the label for position $x$, and $p^\ell$ is the predicted discrete probability distribution from our model.

- The Smooth $l_1$ Loss of Key-point Locations,
  $$L_{pt}^{(2)} = \sum \text{Smooth}_1(F^{(2)}, F^{(3)}),$$  
  where $F^{(2)}$ is the projected 2D key-points calculated according to Eqn 2 from predicted 3D transformation parameters, and $F^{(3)}$ is the ground truth locations.

- The Smooth $l_1$ Loss of Bounding Boxes, $L_{box}^{(5)}$.

The overall loss function is defined by:

$$L = L_{cls} + L_{pt}^{(2)} + L_{box}^{(5)}$$

Experiments

Our method is evaluated on FDDB and AFW. Results are shown in Figure 4 and Figure 5.

Figure 3: Examples of face proposals computed using predicted 3D transformation parameters.

Figure 4: FDDB results based on discrete (left) and continuous scores (right).

Figure 5: Sample qualitative results on the AFW dataset.

Conclusion and Discussion

Our method is a clean and straightforward solution when taking into account a 3D model in face detection, with very compatible state-of-the-art performance obtained.

We are also working on extending the proposed method for other types of rigid/semi-rigid object classes(e.g., cars).

We expect that we will have a unified model for cars and faces which can achieve state-of-the-art performance.
Multi-label Active Learning Based on Maximum Correntropy Criterion: Towards Robustness and Discriminative Labeling

Zengmao Wang¹, Bo Du¹, Lefei Zhang¹, Liangpei Zhang¹, Meng Fang², Dacheng Tao³

Wuhan University¹, University of Melbourne², University of Technology Sydney³

ABSTRACT

Multi-label learning is a challenging problem in computer vision fields. Since annotating a multi-label instance costs greatly, multi-label classification has become a hot topic research. State-of-the-art active learning methods either annotate all the relevant samples without diagnosing discriminative information in the labels or annotate only limited discriminative samples manually, which has weak immunity for the outlier labels. In this study, we focus on the outlier labels to reduce the label information redundancy and the number of labeled instances. A multi-label active learning based on Maximum Correntropy Criterion (MCC) is proposed, which is robust when the outlier happens.

MOTIVATION

In active learning, representative information is important to select the most informative instances. However, it is hard to measure the similarity between the multi-label instances with features when the outlier labels. Figure 3 shows how the outlier labels are influential to the similarity measurement.

To overcome the influence of the outlier labels, we use the MCC to measure the uncertainty and representativeness in multi-label active learning with both label information and features similarity.

THE PROPOSED APPROACH

1. Maximum Correntropy Criterion

Correntropy is a similarity measure between two arbitrary random variables a and b, defined by

\[ V_\sigma(a, b) = E[\phi(a, b)] \]

where \( \phi(a, b) \) is the kernel function that satisfies Mercer theory and \( E[\phi] \) is the expectation operator. With such a definition, the properties of correntropy are symmetric, positive and bounded. Since the joint probability density function of a and b in practice is unknown, and the available data \( (a_i, b_i), i=1, n \) are usually finite, the sample estimator of correntropy is usually adopted by

\[ \hat{V}_\sigma(a, b) = \frac{1}{n} \sum_{i=1}^{n} K_\sigma(a_i, b_i) \]

where the kernel function \( K_\sigma \) is the Gaussian kernel. According to [3], the correntropy between a and b is given by

\[ \max_{a,b} \frac{1}{n} \sum_{i=1}^{n} K_\sigma(a_i, b_i) \]

The objective function is called maximum correntropy criterion (MCC).

2. Uncertainty Measured by MCC

Minimum Margin is a popular approach to measure the uncertainty [4, 5]. We extend it to multi-label learning with MCC loss function, the objective function is as follows:

\[ \Gamma(x_k, y) = \max_{y_k \in \Omega} \left[ E[D(f(x_k), y_k | y)] - \lambda \|f_k\|_2^2 \right] \]

With the similarity measurement, the representative information for each sample is as follows:

\[ \frac{1}{n} \sum_{i=1}^{n} s(x_i, x) \]

Then the representative part can be presented with the whole unlabeled data as follows:

\[ E(U, x_k) - \beta E(U, x_k) = \max_{y_k \in \Omega} \left[ E[D(f(x_k), y_k | y)] - \lambda \|f_k\|_2^2 \right] \]

4. The Objective Function

To utilize the uncertainty and representativeness, we combine the two parts with a tradeoff parameter, and the objective function is as follows:

\[ \max \Gamma(x_k, y) + \beta E(U, x_k) - \beta E(U, x_k) \]

RESULTS

Compared with state-of-the-art methods on 12 multi-label data sets²

CONCLUSION

Contributions

• To the best of our knowledge, it is the first work to focus on the outlier labels in multi-label active learning based on MCC.
• A new approach is derived to make the uncertain information more precise with the prediction labels of the unlabeled data.
• The proposed representative measurement considers labels similarity by MCC. A new way is provided to merge representativeness into uncertainty.

Limitations

• A mechanism should be developed to select the tradeoff parameters adaptively and make the proposed method more practical.

REFERENCES


1http://mulan.sourceforge.net/datasets-mlc.html
Shading-aware Multi-view Stereo
Fabian Langguth, Kalyan Sunkavalli, Sunil Hadap, Michael Goesele

Energy Formulation

Our energy balances geometric errors versus shading errors depending on the local image gradient. This is motivated by Land’s Retinex theory, which assumes that shading introduces only small image gradients, changing the surface brightness gradually. Strong gradients on the other hand are usually caused by changes in surface materials and are thus independent of the illumination.

\[ E(d_i) = \sum_{j,k \in N_i, x_i \in V_i} |E_{g}^{jk}(d_i, x_i)| + \frac{0.01}{\|\nabla I(x_i)\|^2} \]

Geometric Error \(E_g\)
- Multi-view matching using 6–10 neighbors
- Projection \(P\) from main view to neighbor depending on depth \(d\)
- Local gradient based error, identical to [2]

\[ P_j(x_i, d_i(x_i)) = K_j (R R_i^{-1} (K_i^{-1} x_i \cdot d_i(x_i) - t_i) + t_j) \]

\[ E_{g}^{jk}(d_i, x_i) = \nabla I_j(x_i) - \nabla I_j(P_j(x_i, d_i(x_i))) \]

Shading Error \(E_s\)
- Measure the difference between reflectance model \(R\) and image values \(I\)
- Simple Lambertian shading \(S\) with low-frequency spherical harmonics lighting \(L\)

\[ \nabla I(x) - \nabla R(x) \text{ with } R(x) = \alpha(x) \cdot S(n(x), I) \]

Problem: Cannot optimize easily due to unknown albedo \(\alpha\)

\[ \nabla R(x) = \nabla \alpha(x) \cdot S(n(x), I) + \alpha(x) \cdot \nabla S(n(x), I) \]

Solution: Optimize in logarithmic space – Retinex assumption is very beneficial
- If we use the shading error only in regions with very small albedo gradients we can remove the albedo from our energy formulation

\[ \log(R(x)) = \log(\alpha(x)) + \log(S(n(d_i(x)), I)) \]

\[ \nabla \log(R(x)) = \frac{\nabla S(n(d_i(x)), I)}{S(n(d_i(x)), I)} \approx \frac{\nabla S(n(d_i(x)), I)}{S(n(d_i(x)), I)} \]

\[ E_s(d_i, x) = \nabla \log(I(x)) - \nabla \log(R(x)) = \frac{\nabla I(x)}{I(x)} - \frac{\nabla S(n(d_i(x)), I)}{S(n(d_i(x)), I)} \]

Optimization

- Surface is represented as set of bicubic patches [2]
- Patches interpolate depth values and surface normals smoothly over a set of pixels
- Coarse to fine optimization by splitting and adding/removing patches
- Simple Gauss-Newton iterations with reweighting for L1 norm

Results

From top to bottom:
- High amount of detail in Middlebury MVS benchmark, even for sparse input data
- Improved normals compared to standard surface regularization [2]
- Improved geometry compared to shading-based refinement pipeline [1]
- Consistent reconstruction for multi-scale dataset

References

4. Accurate, Dense, and Robust Multi-View Stereopsis, Furukawa, Y., Ponce, J., Transactions on Pattern Analysis and Machine Intelligence 2010
Objective function

- Generative loss: $L_G(Z, Y) = \sum_{i=1}^{N} D_{bce}(G(Z_i), Y_i)$
- Discriminative loss: $L_D(Z, Y) = \sum_{i=1}^{N} D_{bce}(Y_i, 0) + D_{bce}(G(Z_i), 1)$
- Intensity loss: $L_p = \|Y - G(Z)\|_p$
- Angular loss: $L_{ang}(Y, G(Z)) = 1 - Y^T G(Z)$
- Integrability: $L_{curl} = \|\nabla \times G(Z)\|$
Overview

This work presents an open source Multi-View Stereo system for robust and efficient dense modeling from unstructured image collections. Experiments on benchmarks and large-scale Internet photo collections demonstrate state-of-the-art performance in terms of accuracy, completeness, and efficiency.

Contributions

- Joint depth - normal - occlusion inference
  embedded in improved PatchMatch sampling scheme
- Pixelwise view selection
  using photometric and geometric priors
- Multi-view geometric consistency
  for simultaneous refinement and image-based fusion
- Graph-based filtering and fusion
  of depth and normal maps

Joint Depth - Normal - Occlusion Inference

- Joint likelihood function $P(X, Z, \theta, N)$
- Generalized Expectation Maximization (GEM)
  - E-Step: Infer $Z$ using variational inference
  - M-Step: Infer $\theta, N$ using PatchMatch sampling

Multi-View Geometric Consistency

Filtering and Fusion

Results
Laplacian Pyramid Reconstruction and Refinement for Semantic Segmentation

Golnaz Ghiasi, Charless C. Fowlkes
Department of Computer Science, University of California, Irvine

Semantic Segmentation Using CNN

Semantic segmentation assigns a class label to each pixel in the image. Having a large labeled dataset, we can train a CNN to compute a score map for each class label. However, the spatial resolution of high-level feature maps is much lower than the input image due to max pooling and down-sampling.

A common approach to solve this problem is to use bilinear upsampling to compute high-resolution class scores from low-resolution class scores [1]. In this work, we show that better decoding of low-resolution activations and intelligent combination across levels can greatly improve segment fidelity.

Reconstructing High Resolution Segments

- We encode the high-resolution class scores using a bilinear upsampling architecture. Our reconstruction architecture block from the high-dimensional, low-resolution feature maps across levels that limits these tradeoffs.
- The high-resolution class scores from low-resolution class scores [1]. In this work, we show that better decoding of low-resolution activations and intelligent combination across levels can greatly improve segment fidelity.

We use low-resolution feature maps to reconstruct a coarse segmentation and then refine this prediction by adding in higher frequency details derived from higher resolution feature maps.
- Boundary masking (inset) suppresses the contribution of higher resolution layers in areas where the segmentation is confident.
- At each resolution layer, the reconstruction filters perform the same amount of upsampling (e.g., 8x). An additional 2x bilinear upsampling is then applied to each class score map before combining it with higher resolution predictions.

Experimental Performance Evaluation

- Base model is trained using 11k training images from PASCAL VOC 2011 augmented with multiple scaled versions (0.5x-1.5x) cropped to a max dimension of 384. Final model is trained with 97k additional images from MS COCO dataset.

- Adding in higher resolution feature maps results in the most performance gain near object boundaries while masking improves performance both near and far from boundaries.
- Running the model at multiple scales (ms) and post-processing with a conditional random field (crf) yields further small improvements.
- We compare our model without masking to the baseline FCN model [1] which uses bilinear upsampling on PASCAL VOC 2011 validation data.

- Our model achieves mean IoU of 79.3% on PASCAL VOC 2012 test data and 71.8% on CityScape test data which are among top performing model submitted to the test benchmark servers.

References


The models and code are available at https://github.com/golnazghiasi/LRR
Generic 3D Representation via Pose Estimation and Matching

Amir R. Zamir  Tilman Wekel  Pulkit Agrawal  Jitendra Malik  Silvio Savarese

- Fully Supervised Learning is task-specific.
- Lacks generalization and abstraction.
- Cannot lead to a comprehensive human-like vision.
- **Proposal:** supervision over a set of selected foundational tasks → generalization to unseen tasks and abstraction capabilities.
- **Foundational Tasks:** Inspiration From Biology.
- **Goal:** A generic (3D) representation → no fine-tuning needed.

**Inspiration from Biology:**
- Developmental Stages of Human Vision → Foundational Tasks

**Learning Framework:**

**Dataset** (automatically collected from Street View)

**Sample Collected Data**

**Evaluations** (Camera Pose Estimation & Wide Baseline Matching)

**Unsupervised Evaluations** (3D Scene Layout, Surface Normals, 3D Object Pose Estimation)

**Under the hood:** vanishing points?

Demo, Data, Code, Results: http://3drepresentation.stanford.edu/
### 1 Highlights

- Hand pose estimation from single depth image with surface normal
- Local reference frames with local surface normals are invariant to rigid body transformation, so no data augmentation is needed
- Local surface normal difference is invariant to rigid body transformation and better captures the local geometrical property of hand surface, e.g., curvature

### 2 Random normal difference

**Random normal difference feature**

- $\delta (\cdot, \cdot)$ is calculated with the help of local surface normal, invariant to in plane rotation:
  - for edge point, determined by normal
  - for inner point, determined by normal + reference point

- $\Delta (\cdot, \cdot)$ is the dot product of two surface normal directions, invariant to rigid body transformation, while depth difference is not invariant to rigid body transformation.

**Calculating random position**

$f_{T}(p_i, \delta_1, \delta_2) = \Delta(\phi_T(r(p_i, \delta_1)), \phi_T(r(p_i, \delta_2)))$

### 3 Random normal difference

**Invariance to rigid body transformation**

- Random offset invariant to rigid body transformation, i.e., the relative position between $p_i$ and $r(p_i, \delta)$ remains unchanged after transformation:
  \[
  T(p_i - r(p_i, \delta)) = T(p_i) - T(r(p_i, \delta))
  \]

- Feature channel difference is invariant to rigid body transformation:
  \[
  \Delta(\phi_T(q_1), \phi_T(q_2)) = \Delta(\phi_T(q'_1), \phi_T(q'_2))
  \]

### 4 Hierarchical estimation

**Frame Conditioned Regression Forest**

- Regress the offset between joint $j$ and point $i$ conditioned on referend frame
- Training: offset is normalized w.r.t. the referenced frame
- Testing: firstly regress w.r.t. referenced frame, then transformed into the global frame

**Palm Pose Estimation**

- wrist location estimated with only edge points
- surface normal assumed orthogonal to the image plane based on orthographic projection

**Finger Pose Estimation**

- use the palm frame to estimate the PIP points
- use the finger frame to estimate the DIP/TIP points
- More details can be found in the paper

### 5 Results

**References:**

Abundant Inverse Regression using Sufficient Reduction and its Applications


http://cs.wisc.edu/~hwkim/projects/air/

**OBJECTIVE**

**Goal:** Develop a regression model explaining why a particular prediction was made at the level of each example/sample

**Strategy:** Inverse Regression and Sufficient Reduction in the “abundant” feature setting.

**MAIN IDEA**

Desired: Relevance of individual covariates at the level of specific samples for a given regression task.

Challenge (“Chicken-or-egg problem”): Relevance/confidence score to individual covariates $x^i$ should condition the estimate based on knowledge of all other (uncorrelated) covariates $x^j$.

**Solution:** sufficient reduction

$f(x^i | \phi(x))$, where $x^i | X, \phi(X) \sim x^i | \phi(X)$

Desired: Robust regression model which allows missing or randomly corrupted covariates with their dynamic weights.

**Algorithm**

1. **Training**
   - Estimate a joint distribution for each covariate, $f(x^i, y)$
   - Find sufficient reduction $\phi : x^i \rightarrow y$ for each subset of features $x^i$

2. **Estimate the prior/weight for $\phi()$ as $w_{\phi} = \mathbb{E}[(y - \phi(x^i))^2]^{-1}$**

3. **Estimate cond. confidence of feature $w_{\phi^j} = \sum_i w_i f(x^i | y^i)/ \sum_i w_i$**

4. **Fit a feature confidence aware regressor $h : \{[x^i]_{x^i < 1}, \{w_i \}_{i=1} \rightarrow y$**

5. **Prediction**
   - Evaluate $w_{\phi} := \mathbb{E}f(x^i | \phi(x^i))$ by lines 3 and 5, with learned $w_{\phi}$.

   $\hat{y} = h(x^i | \phi(x^i))$

**LEMMA 1: OPTIMAL GLOBAL WEIGHS FOR $\phi^j$**

Suppose we have $K$ random variables (sufficient reduction),

$\phi^i(x_k) \sim N(y, \sigma_1^2), \ldots, \phi^K(x_K) \sim N(y, \sigma^2_K)$,

where $\sigma_1^2 > 0, \forall k \in \{1, \ldots, K\}$. Consider a convex combination of $\phi^j$. Its expectation is $y$.

Assuming that the errors of all sufficient reductions are independent, the problem to find the optimal weights for the convex combination with the minimal variance can be formulated as

$$\min_w \sum_{k=1}^{K} \sigma_k^2(w_k) \text{ s.t.} |w_k| = 1 \text{ and } w_k \geq 0, \forall k \in \{1, \ldots, K\}$$

The unique global optimum of Eq. (4) is $w_k = \sigma_k^{-2} / \sum_{k=1}^{K} \sigma_k^2$.

**EXPERIMENTS: AMBIENT TEMPERATURE PREDICTION**

![Figure](image.png)

**ALGORITHM**

1. **Training**
   - Estimate a joint distribution for each covariate, $f(x^i, y)$
   - Find sufficient reduction $\phi : x^i \rightarrow y$ for each subset of features $x^i$

2. **Estimate the prior/weight for $\phi()$ as $w_{\phi} = \mathbb{E}[(y - \phi(x^i))^2]^{-1}$**

3. **Estimate cond. confidence of feature $w_{\phi^j} = \sum_i w_i f(x^i | y^i)/ \sum_i w_i$**

4. **Fit a feature confidence aware regressor $h : \{[x^i]_{x^i < 1}, \{w_i \}_{i=1} \rightarrow y$**

5. **Prediction**
   - Evaluate $w_{\phi} := \mathbb{E}f(x^i | \phi(x^i))$ by lines 3 and 5, with learned $w_{\phi}$.

   $\hat{y} = h(x^i | \phi(x^i))$
Microlens arrays are sheets of plastic that have a grid of small lenses. These lenslets focus parallel rays of light onto a pattern to create different appearances per viewpoint. We create a pattern from black squares randomly placed on a white background. This makes each lenslet have an independent black and white appearance and we only consider a discrete appearance. Therefore, we encode viewpoint by considering the appearance of all lenslets together.

For a single lenslet, we record its discrete appearance for a set of viewpoints. Using this calibration for each lenslet, we can infer the viewpoint.

We design and demonstrate a passive physical object whose appearance changes to give a discrete encoding of its pose. This object is created with a microlens array that is placed on top of a black and white pattern. When viewed from a particular viewpoint, the lenses appear black or white depending on the part of the pattern that each microlens projects towards that viewpoint. We introduce the process through which the discrete microlens pattern can be turned into a viewpoint and a pose estimate and evaluate viewpoint and pose estimation accuracy.

### Number of Lenslets

<table>
<thead>
<tr>
<th>Number of Lenslets</th>
<th>Error of Random Viewpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1.2</td>
</tr>
<tr>
<td>30</td>
<td>1.5</td>
</tr>
<tr>
<td>40</td>
<td>1.8</td>
</tr>
</tbody>
</table>

### Lighting Environment

<table>
<thead>
<tr>
<th>Light Source</th>
<th>Error of Viewpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overhead</td>
<td>1.2</td>
</tr>
<tr>
<td>Spotlight</td>
<td>1.5</td>
</tr>
<tr>
<td>Starlight</td>
<td>1.8</td>
</tr>
</tbody>
</table>

### Rotation Estimation Results

<table>
<thead>
<tr>
<th>Lighting Environment</th>
<th>Traditional Marker</th>
<th>Microlens Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dot Rotation</td>
<td>1.2035</td>
<td>0.5405</td>
</tr>
<tr>
<td>Estimation Error</td>
<td>0.9373</td>
<td>0.50538</td>
</tr>
<tr>
<td>z=0.015</td>
<td>1.5076</td>
<td>0.56297</td>
</tr>
</tbody>
</table>

Our algorithm considers the statemap of each lenslet... and votes for the viewpoints that are consistent with each lenslet appearance. The viewpoint with the most votes is the inferred viewpoint.
1. Action Understanding

- What happens in a video? (Video classification)
- When does it happen? (Action Temporal Localization)

Goal: Efficiently retrieve temporal segments from untrimmed videos which are likely to contain human actions.

Motivation:
- Video data is inherently untrimmed.
- Delimiting when actions occur is the first stage for rich video understanding.

2. Temporal Action Proposals

- Action proposals -> efficient approaches to retrieve temporal segments likely to contain actions.
- This work leverages LSTM memory networks to encode multiple temporal scales instead of exhaustively analyze them.

3. DAPs Network

Inference
- Slide the recurrent network over streams of length T (512 frames) over the video with stride and collect all the outputs after evaluating each stream

Learning
- Relaxation of an optimal assignment problem:

\[
(x^*, \theta^*) = \arg \min_{x, \theta} \alpha \mathcal{L}_\text{match}(x, S(v, \theta), A(v)) + \mathcal{L}_\text{conf}(x, C(v, \theta))
\]

s.t. \( x_{ij} \in \{0, 1\}, \sum_{i} x_{ij} = 1 \)
- Match to K anchors, summarizing location and duration of actions, instead of \( A(v) \)
- Penalize predictions far from anchors that match a ground-truth
- \( \mathcal{L}_\text{match}, \mathcal{L}_\text{conf} \) are \( l_2 \) and Cross-Entropy loss, respectively

4. Recall Analysis in THUMOS-14

- THUMOS-14 -> More than 24 hours of labeled videos with 20 sports categories
- Metric -> Average Recall for a range of tIoU values as a function of the number of proposals.
- Outperform multi-scale supervised methods with hand-crafted or “deep” representations as well as unsupervised approaches

5. Efficient Run-time

- Efficient scanning by avoiding multi-scale exploration
- Power-up by GPU parallel processing

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Feature</th>
<th>Time [seconds]</th>
<th>Speedup</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>APT</td>
<td>-</td>
<td>285.8</td>
<td>1.0</td>
<td>0.68</td>
</tr>
<tr>
<td>BoFrag</td>
<td>312.3</td>
<td>794.8</td>
<td>1.0</td>
<td>0.68</td>
</tr>
<tr>
<td>Sparse-prop</td>
<td>191.1</td>
<td>533.6</td>
<td>14.9</td>
<td>10.2</td>
</tr>
<tr>
<td>SCNN-prop</td>
<td>N.A</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DAPs</td>
<td>N.A</td>
<td>1.34</td>
<td>5931.9</td>
<td>134.1</td>
</tr>
</tbody>
</table>

6. Qualitative Results

- Top ranked proposals
- Matched proposals vs. temporal relevance

7. Does it work for unseen actions?

- Trained on 20 sports categories (THUMOS-14) and tested on a richer diverse set of activities, ActivityNet

8. Action Detection

- Assess what and when action occurs in a video
- Evaluation in THUMOS-14 in terms of Average Precision (Precision & Recall)
- Promising results for action detection

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>Method</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>13.9</td>
<td>Caba Heilbron et. al [4]</td>
<td>13.2</td>
</tr>
</tbody>
</table>

9. Conclusions

- It is possible to leverage memory cells to learn an appropriate encoding of the video sequence which compacts multiple action durations on a single scale
- Promising effective and efficient architecture for temporal action localization
- Future work -> end-to-end representation and multi-task learning approaches
Introduction

We have created a large diverse set of cars from overhead images, which are useful for training a deep learner to binary classify, detect and count them. The dataset and all related material will be made publicly available. To contain contextual information in its training, we demonstrate classification and detection on this dataset using a neural network we call Resception. This network combines residual learning with Inception-style layers and is used to count cars in one look. This is a new way to count objects rather than by localization or density estimation. It is fairly accurate, fast and easy to implement. Additionally, the counting method is not car or scene specific. It would be easy to train this method to count other kinds of objects and counting over new scenes requires no extra setup or assumptions about object locations.

Network and Context

Test patches the network correctly classified as containing a car in the central region. Occlusions and visibility issues are commonly handled, but we note that they still appear to account for much of the error. The left most image is not a mistake. It has a tree in the center while the shifted version above it has a car concealed slightly underneath the tree.

Detection

We tested detection on a held out set of 10 images comprising approximately 1 square km. Simple non-maximal suppression is used.

| Condition | Count | TP | FP | TN | Precision | Recall | F  
|-----------|-------|----|----|----|-----------|--------|---
| Verification | 260 | 253 | 7 | 20 | 96.04% | 97.31% | 96.65% |
| Detection | 260 | 250 | 10 | 10 | 92.59% | 96.15% | 94.35% |

Below can be seen several examples of detection on many scenes. Three of these are new scenes from post publication.

Fast Counting

Examples of patches which were correctly counted by the network. From left to right the correct number is 9, 3, 6, 13 and 47. Note that cars which are not mostly inside a patch are not counted. The center of the car must be at least 8 pixels inside the visible region.

Fast counting takes in 224x224 patches in each stride which is 167 pixels. These are results on 20 test scenes like the ones above. The color borders show the extent of each patch/stride. The numbers in red are the count for each patch.

| Model | NMS | Time | Stride | Count | TP | FP | TN | Precision | Recall | F  
|-------|-----|------|--------|-------|----|----|----|-----------|--------|---
| AlgoNet | 8.40 | 8.40 | 128 | 27 | 20 | 20 | 97.35% | 96.04% | 96.65% |
| Inception | 8.40 | 8.40 | 128 | 27 | 20 | 20 | 97.35% | 96.04% | 96.65% |
| Resception | 8.40 | 8.40 | 128 | 27 | 20 | 20 | 97.35% | 96.04% | 96.65% |

Large strides allow one to count very fast. Using a single Titan X (Maxwell) we can count the cars in as fast as 1 square kilometer per second. This can easily be parallelised to count even faster than that.

http://gdo-datasci.ucllnl.org/cowc/
Reliable Attribute-Based Object Recognition Using High Predictive Value Classifiers
Wentao Luan\textsuperscript{1}, Yezhou Yang\textsuperscript{2}, Cornelia Fermüller\textsuperscript{2}, John S. Baras\textsuperscript{1}
\textsuperscript{1} Institute for Systems Research, \textsuperscript{2} Computer Vision Lab, University of Maryland, College Park

Introduction

Goal: Reliable 3D object recognition using an ensemble of attribute-based classifiers for active agents.

Motivation:
- Insufficient representative training samples make it difficult to learn the optimal positive and negative detection rate.
- The viewing conditions can have a strong influence on classification performance.

Methods and Intuition:
- Classifiers use two thresholds with one aiming for a positive predictive value (PPV), giving high precision for positive classes, and the other aiming for a negative predictive value (NPV), giving high precision for negative classes.
- High PPV and NPV thresholds should be easier to obtain when the number of training samples is too small to represent well the underlying distribution.
- Incorporating environmental factors (distance) into decision making. Defining a reliable working region for each basic attribute classifier, indicating a fair separation of the distributions of positive and negative classes.
- Hence our approach can actively select “safe” samples and discard “unsafe” ones in unreliable regions.
- Allows using simple attributes.
  - Simple attributes are usually more robust to viewing conditions, though less discriminative.

Framework

Learning phase (offline): We learn distance-dependent attribute classifiers and determine a reliable range of distance intervals for each attribute classifier.

Testing (online): The system decides the distance interval for the RGBD images from active agents; combines classifier measurements from multiple images via maximum a posteriori probability (MAP) estimation. As illustrated in Figure 1 (right),

$$P(O = o_j | Z^k, E^k) = \lambda P(O = o_j) \prod_{k=1}^{K} \prod_{i=1}^{M} \frac{P(F_i = f_{ij} | Z^k, E^k)}{1 - \sum_{i=1}^{M} P(O = o_i)}.$$ 

The recognition $A$ is derived using MAP estimation as:

$$A = \arg\max P(O = o_j | Z^k, E^k).$$

Only takes high predictive value observations in reliable working region:

$$P(F_i = 1 | Z^k, E^k) = \begin{cases} p_i, & \text{if } e_k \in R_i \text{ & } z^k_i = 1, \\ 1 - p_i, & \text{if } e_k \in R_i \text{ & } z^k_i = 0, \\ \sum_{i=1}^{M} P(O = o_i) & \text{o.w.} \end{cases}$$

$R_i$ is the set of reliable working regions for the $i$-th classifier.

Experimental Validation

Set of 9 objects and 10 attributes in the range of 0.6-1.6m

Attributes:
- Fine shape: 1) retrieve templates point cloud based on input’s location w.r.t. the camera 2) match the templates using VFH features
- Coarse shape: plane and cylinder fitting using RANSAC
- Color: histogram matching in hue and saturation space

Figure 2: Object IDs and their list of attributes. The dataset is available from http://ece.umd.edu/~wluan/ECCV2016.html

Figure 3: Illustration of the preprocessing pipeline. Left: input point cloud; Middle: point cloud after pass-through filtering; Right: segmented candidate object and removed table surface.

Experiment One: demonstrates the necessity of incorporating environmental factors (the recognition distance in our case) for object recognition

Experiment Two: evaluates the performance of the high predictive value threshold classifier in comparison to the single threshold one

Figure 4: Estimated distribution of bottle shape classifier’s response score under 4 recognition distance intervals. The classifiers’ response score distributions are indeed distance variant.

Figure 5: Left: error rate using classification with a single threshold (blue) and two high predictive value thresholds (red). The green line depicts the error component due to the cases where the two thresholds method picks randomly. Right: three systems’ recognition accuracy for different working distance intervals.

Figure 6: Component classifiers’ and their combinations’ response scores for different recognition distance intervals.

Figure 7: Our method provides an alternative for cases when the training data in real world scenarios does not represent well the underlying distribution.

Experiment Three: demonstrates the benefits of using less discriminative attributes for extending the system’s working range.

- The classification accuracy decreases with larger distances.
- At 120 cm to 160 cm, the system using fine shape attributes (blue) performs worse than the system using less selective coarse attributes (green). It validates that the coarse shape based classifier has a larger working region.
- The system using all attributes (yellow) achieves the best performance at each working region.

Our method provides an alternative for cases when the training data in real world scenarios does not represent well the underlying distribution.

Experiment Three: demonstrates the benefits of using less discriminative attributes for extending the system’s working range.

- The classification accuracy decreases with larger distances.
- At 120 cm to 160 cm, the system using fine shape attributes (blue) performs worse than the system using less selective coarse attributes (green). It validates that the coarse shape based classifier has a larger working region.
- The system using all attributes (yellow) achieves the best performance at each working region.

Theorem 1: The system is guaranteed to provide correct recognition if
- The recognized attributes can differentiate an object from others;
- The component classifiers’ predictive values are larger than specified values (details see paper).

Theorem 2: The MAP estimation will converge to the correct result if
- The attribute classifiers’ PPV and NPV are high enough in their reliable working region, where a lower bound of detection rate exists;
- The inputs are sampled randomly such that each attribute classifier gets the same chance to work in its reliable region.
Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition
Jun Liu¹  Amir Shahroudy¹  Dong Xu²  Gang Wang¹
¹Nanyang Technological University  ²University of Sydney

Contributions
- Spatio-temporal design of LSTM network for 3D action recognition.
- Skeleton-based tree traversal technique to feed the structure of the skeleton data into a sequential LSTM.
- Improving the design of spatio-temporal LSTM by adding the trust gate to deal with noisy input.

1 Spatio-Temporal LSTM
- Motivation: (1) LSTM network is suitable for modelling dependence in temporal domain. (2) There is also high dependence among joints in spatial domain.
  - Spatio-temporal LSTM.
    - Spatial direction: body joints are fed in a sequence.
    - Temporal direction: locations of corresponding joints are fed frame by frame.

2 Tree-Structure based Traversal
- It is beneficial to model the spatial dependency of the joints based on their adjacency tree structure (see figure c)

3 Spatio-Temporal LSTM with Trust Gate
- Unreliable input data (noisy joint coordinates) restricts classification accuracy.
- Learn to predict the input using contextual information:
  \[ p_{j,t} = \tanh \left( M_p \left( h_{j-1,t}, h_{j,t-1} \right) \right) \]
  Prediction of input
  Contextual information
- Assess reliability of actual input by comparing it with the prediction:
  \[ x'_{j,t} = \tanh \left( M_x \left( x_{j,t} \right) \right) \]
  Actual input
  \[ \tau_{j,t} = G(x'_{j,t} - p_{j,t}) \]
  \[ G(z) = \exp(-\lambda z^2) \]
- Updating cell state using trust gate \( \tau_{j,t} \):
  \[ c_{j,t} = \tau_{j,t} \odot i_{j,t} \odot u_{j,t} + (1 - \tau_{j,t}) \odot f_{j,t} \odot c_{j-1,t} + (1 - \tau_{j,t}) \odot i_{j,t} \odot c_{j-1,t-1} \]
- Trustable: Update memory cell by importing new input information.
- Unreliable: Take more history information and try to block new input.

4 Action Recognition Results
- NTU RGB+D
- SBU Interaction
- UT-Kinect

Contributions
- Spatio-temporal design of LSTM network for 3D action recognition.
- Skeleton-based tree traversal technique to feed the structure of the skeleton data into a sequential LSTM.
- Improving the design of spatio-temporal LSTM by adding the trust gate to deal with noisy input.

Visualization of Trust Gate
- MSR3D dataset
- MHAD motion capture dataset

\( j_1, j_2, j_3 \): Accurate joint
\( j_3' \): Noisy joint
\( (i_N, t_N) \): Noisy spatio-temporal step
1. Introduction

- **Motivation**
  - Video recognition usually requires a large number of training examples, which are expensive to be collected.
  - An alternative and cheaper solution is to draw from the large-scale images and videos from the Web.
  - With modern search engines, the top ranked images and videos are usually highly correlated to the query.

- **Challenges**
  - Web images and video frames are typically noisy and may be of completely different domains from that of users’ interests (e.g., cartoons vs. natural images).
  - Web videos are usually untrimmed and very lengthy, where some query-relevant frames are often hidden in between the irrelevant ones.

2. Key Observations

- The relevant images and video frames typically exhibit similar appearances, while the irrelevant images and videos have their own distinctiveness.
- Selecting training examples from Web images and videos can be made easier, if they could be mutually filtered to keep those in common.

3. Approach

We first jointly choose images and video frames and try to match them aggressively, and then impose a passive constraint over the selected video frames, such that the frames are not too far from the original videos.

\[
\min_{\alpha \in [0, 1]^M, \beta \in [0, 1]^W} \left( \begin{pmatrix} \alpha^T & \beta^T \end{pmatrix} \begin{pmatrix} K_I & -K_{IV} \\ -K_{IV}^T & K_V \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \lambda \| V - V \cdot \text{diag}(\beta) \cdot W \|_2^2 \right) \]

4. Experiment

- **Action recognition on UCF101**

<table>
<thead>
<tr>
<th>Method</th>
<th># Number of training data</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All crawled data</td>
<td>426K</td>
<td>64.7</td>
</tr>
<tr>
<td>Validation</td>
<td>368K</td>
<td>66.5</td>
</tr>
<tr>
<td>One-class SVM (10%)</td>
<td>384K</td>
<td>65.9</td>
</tr>
<tr>
<td>One-class SVM (15%)</td>
<td>363K</td>
<td>65.9</td>
</tr>
<tr>
<td>Unsupervised One-class SVM (10%)</td>
<td>384K</td>
<td>66.4</td>
</tr>
<tr>
<td>Unsupervised One-class SVM (15%)</td>
<td>363K</td>
<td>66.4</td>
</tr>
<tr>
<td>Landmarks (10%)</td>
<td>364K</td>
<td>67.7</td>
</tr>
<tr>
<td>Landmarks (15%)</td>
<td>363K</td>
<td>68.3</td>
</tr>
<tr>
<td>Ours (10%)</td>
<td>384K</td>
<td>69.3</td>
</tr>
<tr>
<td>Ours (15%)</td>
<td>363K</td>
<td>68.9</td>
</tr>
</tbody>
</table>

- **Video event detection on TRECVID MED 2013**

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Composite concept</td>
<td>6.4</td>
</tr>
<tr>
<td>EventNet</td>
<td>8.9</td>
</tr>
<tr>
<td>Selecting</td>
<td>11.8</td>
</tr>
<tr>
<td>Ours</td>
<td>16.1</td>
</tr>
</tbody>
</table>

5. Conclusion

- We investigated to what extent Web images and videos could be leveraged jointly to conduct Webly-supervised video recognition.
- We expect this work to benefit future research on large-scale video recognition tasks.

Acknowledgement: This work was supported in part by NSF IIS-1566511.
In this paper, we propose a real-time system, Hierarchical Feature Selection (HFS), that performs image segmentation at a speed of 50 frames-per-second. We make an attempt to improve the performance of previous image segmentation systems by focusing on two aspects: (1) a careful system implementation on modern GPUs for efficient feature computation; and (2) an effective hierarchical feature selection and fusion strategy with learning. Compared with classic segmentation algorithms, our system demonstrates its particular advantage in speed, with comparable results in segmentation quality. Adopting HFS in applications like salient object detection and object proposal generation results in a significant performance boost. Our proposed HFS system (will be open-sourced) can be used in a variety computer vision tasks that are built on top of image segmentation and superpixel extraction.

**Abstract**

Hierarchical Feature Selection (HFS)

**Workflow**

**Sample Results**

**Applications**


**References**

Ming-Ming Cheng1, Yun Liu1, Qibin Hou1, Jiawang Bian1, Philip Torr2, Shi-Min Hu3, Zhuowen Tu4
1ICCE&CS, Nankai University 2Oxford University 3Tsinghua University 4UCSD

**HFS: Hierarchical Feature Selection for Efficient Image Segmentation**

**Boundary Evaluation**

**Region Evaluation**

**Sample Results**

http://mmcheng.net/hfs/
Motivation
- Explanations are important for understanding and interacting with intelligent systems.

Visual Explanations
- Visual explanations must be image relevant and class consistent.
- Justification vs. Introspection: explanations detail why a prediction is compatible with visual evidence.
- Introspective explanations detail the internal mechanisms of a model.
- Here, concentrate on justifications as they are useful for non-experts.
- Method:
  - Introduce novel discriminative loss based on REINFORCE [1].

Data
- Use fine-grained descriptions collected in [2].
- 200 bird classes, > 11k images, 5 captions/image.
- Fine-grained descriptions more class informative than attributes.

Metrics
- Measure image relevance using CIDEr.
- Measure class relevance using class similarity rank.

Relevance Loss
Encourages sentences to be image relevant by minimizing cross entropy between ground truth and predicted words.

\[
L_R = \sum_{t=1}^{T-1} \log p(w_{t+1} | w_{0:t}, I, C)
\]

where \( I \) is an image feature, \( w_t \) is a ground truth word, and \( C \) is a class label.

Discriminative Loss
Encourages sentences to be class discriminative by maximizing a reward, \( R_D \), which measures class discriminativeness of a sampled sentence \( \hat{w} \).

\[
L_D = -E_{\hat{w} \sim p(\hat{w})} [R_D (\hat{w})]
\]

Estimate \( L_D \) with Monte Carlo sampling because computing an expectation over descriptions is intractable.

Use REINFORCE [1] to estimate the expected gradient of \( L_D \):

\[
\nabla_{E_{\hat{w} \sim p(\hat{w})} [R_D (\hat{w})]} E_{\hat{w} \sim p(\hat{w})} [R_D (\hat{w})E_{\hat{w}} \log p(\hat{w})]
\]

Results
- Explanations are more class relevant than descriptions.
  - This is a White Necked Raven because...
    - Description: ...this bird is black in color with a black head and black eye rings.
  - This is a Common Raven because...
    - Description: ...this bird is black in color with a long black tail feather and pointy black eye.
  - This is a Brown Pelican because...
    - Description: ...this bird has a long neck and long bill.
  - This is a Scarlet Tanger because...
    - Description: ...this is a red crow, a short bill, and a red belly.
  - This is a Cardinal because...
    - Description: ...this bird has a red head, a short bill, and a red breast.

- Explanations are more image relevant than definitions.
  - This is a Crested Auklet because...
    - Definition: ...this bird is black with a bright orange bill and white eye.
  - This is a Red Headed Woodpecker because...
    - Definition: ...this bird has a red crown, a white belly, and a white wing bar.
  - This is a Ruby Throated Hummingbird because...
    - Definition: ...this bird is white and brown in color with a long skinny beak, and brown eye rings.

- Explanations for incorrect classification decisions.
  - Correct: Anna Hummingbird
    - Predicted: Green Violetear
  - Correct: Anna Hummingbird
    - Predicted: Anna Hummingbird

Quantitative Evaluations
- Human Evaluations
  - Expert Rank: Lower is better
  - Discriminativeness: Higher is better

Acknowlegments
- This work was funded by Facebook Reality Labs.

References

Lisa Anne Hendricks¹, Zeynep Akata², Marcus Rohrbach¹,
Jeff Donahue¹, Bernt Schiele², Trevor Darrell²
¹University of California Berkeley, ²Max Planck Institute for Informatics
The recovery of 3D human pose with monocular camera is an inherently ill-posed problem due to the large number of possible projections from the same 2D image to 3D space. Human observers are able to accurately estimate the pose of a human body with a single eye by leveraging vast memories of anatomical structure of human body. The difference of the recovered human poses in consecutive frames should be consistent with the actual velocity of each joint.

**Contributions**
- The RGB image and its calculated height-map are combined to detect the landmarks of 2D joints with a dual-stream ConvNet.
- We formulate a new objective function to estimate 3D motion from the detected 2D joints in the monocular image sequence, which reinforces the temporal coherence constraints on both the camera and 3D poses.

**Motivation**
- Image $\rightarrow$ 2D: minimize the error over the locations $l$ and types $t$ of joints
  $$F(l, t) = \sum_{i,j \in E} \theta(t) + \sum_{l \in U} \left| R(l, t, \theta) - \hat{w}(l) \right|$$

**Framework**
- Training of the dual-stream ConvNet: The RGB stream is pre-trained on LSP dataset, and the resultant network is further applied on our synthetic height-maps dataset to obtain the initial weights of the height stream. The entire network is then jointly fine-tuned on a target training set.

**Evaluation of 2D Joints Localization**

**Evaluation of 3D Motion Recovery with Predicted 2D Joints**

### Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Walking</th>
<th>Jogging</th>
</tr>
</thead>
<tbody>
<tr>
<td>HumanEva</td>
<td>130.71</td>
<td>118.69</td>
</tr>
<tr>
<td>Human3.6M</td>
<td>132.37</td>
<td>121.56</td>
</tr>
<tr>
<td>Ours (3 actions)</td>
<td>104.90</td>
<td>121.56</td>
</tr>
<tr>
<td>Ours (15 actions)</td>
<td>117.08</td>
<td>121.56</td>
</tr>
</tbody>
</table>

### Evaluation of 3D Motion Recovery with Ground-Truth 2D Joints

**References**

We define RBF kernels on 3D joint sequences that can compactly capture higher-order relationships between skeleton joints for 3D action recognition.

We embed low-level 3D joints and time index into Reproducing Kernel Hilbert Space.

We devise a sequence compatibility kernel that captures the spatio-temporal compatibility of joints in one sequence against those in the other sequence.

We devise a dynamics compatibility kernel that explicitly models the action dynamics of a sequence, e.g., captures displacement vectors between body joints across sequences.

Evaluating kernels is costly. Thus, we linearize the kernels to form kernel descriptors. The higher-order outer-products derived from these kernel descriptors become higher-order tensor representations.

Kernels $G_{ij}$ capture sensor uncertainty in body-keypoint detection (we use a delta function).

Kernels $G_{ij}$ models the spatio-temporal co-occurrences of the body-joints.

Kernels $G_{ij}$ encode the temporal start and end-points $(s, s')$ from $\Pi_A$ and $(t, t')$ from $\Pi_B$.

Kernels $G_{ij}$ limits contributions of dynamics between temporal points ($\sigma_r$ is small) if they are distant from each other, i.e., if $s > s'$ or $t > t'$.

Our final representation can be expressed as follows:

$$K^*_D(\Pi_A, \Pi_B) = \frac{1}{N} \sum_{(s, t) \in J} \sum_{(s', t') \in J} G_{s't'}((x_{s1} - x_{t1}) - (y_{s1} - y_{t1})) G_{s't'}(s-s') G_{s't'}(t-t').$$

where $G_{s't'}(s-s') = \frac{1}{N} G_{s't'}(s-s').$

Our final representation can be expressed as follows:

$$K^*_S(\Pi_A, \Pi_B) = \frac{1}{N} \sum_{(s, t) \in J} \sum_{(s', t') \in J} G_{s't'}((x_{s1} - x_{t1}) - (y_{s1} - y_{t1})) G_{s't'}(s-s') G_{s't'}(t-t').$$

where $\Lambda$ is a normalization constant, $I = I_A \times I_B$ and $r$ is the order of the statistics, e.g., $r \geq 3.$

Kernels $G_{ij}$, $G_{ij}$, $G_{ij}$ capture the compatibility between (i) joint-types $i$ and $j$, (ii) joint locations $x$ and $y$, and (iii) the temporal alignment of two poses.

Then, we obtain a linearized representation $K^*_S(\Pi_A, \Pi_B) = (\mathbf{V}, \mathbf{V}),$ where:

$$\mathbf{V} = [\mathcal{G}(\mathbf{X})]_{(i,j) \in I}, \text{ and } \mathbf{X} = \frac{1}{\sqrt{N}} \sum_{s \in S} \mathbf{G}(\mathbf{x}_s).$$

Eigenvalue Power Normalisation $\mathbf{G}$ deals with correlated bursts in each sequence:

$$\mathcal{E}(\mathbf{X}) = \text{Sgn}(\mathbf{X}) \mathbf{E} \mathbf{X}.$$
1. Motivation

Given a point cloud of a Manhattan-world building, we want to approximate its geometry using a set of boxes.

**Manhattan-world assumption:** the major components of a building consist of axis-aligned planes.

2. Method

2.1 Candidate box generation

- RANSAC-based plane extraction [Schnabel et al. 2007].
- Iterative plane refinement.
- Detecting missing walls and roofs.
- Partitioning the space into a non-uniform grid.

2.2 Energy terms

- **Data fitting** measures how well a candidate box is supported by the point cloud.

\[
S(b_i) = \sum_j \sum_{f_i} n_f \cdot n_j \cdot \text{dist}(p_j)
\]

\[
\text{dist}(p_j) = \begin{cases} 
1/(t + d)_i, & d_i < d_i \\
0, & \text{otherwise}
\end{cases}
\]

Three types of candidate boxes:
- Positive boxes: positive scores, inside the building (blue).
- Negative boxes: negative scores, outside the building (green).
- Blank boxes: scores close to zero (white).

- **Compactness** discourages holes and protrusions, defined on adjacent boxes:

\[
C_{i,j} = \frac{1}{\min(t_i, t_j)}
\]

i and j are the thickness values of two adjacent boxes \(b_i\) and \(b_j\).

2.3 Optimization

- **A Markov Random Field formulation.** The nodes of the graph represent all candidate boxes and edges connect adjacent boxes.

\[
E(X) = \sum_{b_i} D(b_i) + \lambda \cdot \sum_{(b_i, b_j) \in E} V(b_i, b_j)
\]

where \(D(b_i) = \begin{cases} -S(b_i), & \text{for positive boxes} \\
S(b_i), & \text{otherwise}
\end{cases}\)

and \(V(b_i, b_j) = \begin{cases} C_{i,j}, & \text{if } \min(t_i, t_j) \leq 1 \\
1, & \text{otherwise}
\end{cases}\)

- **Data term** \(D(b_i)\) encourages to choose boxes having higher data fitting scores.
- **Smoothness term** \(V(b_i, b_j)\) discourages holes and protrusions.

3. Results

**Reconstruction from airborne LiDAR data**

**Reconstruction from MVS point clouds**

**Comparisons.** From left to right: MVS points, 2.5D DC [Zhou et al. ECCV’10], L1-based polycube [Huang, et al. TOG’14], and our result.
From Multiview Image Curves to 3D Drawings

Anil Usumezebas\textsuperscript{11}, Ricardo Fabbri\textsuperscript{1} and Benjamin Kimia\textsuperscript{1}
Brown University ~ Rio de Janeiro State University ~ SRI International

**MOTIVATION**

From a large sequence of 2D images, produce 3D models

This paper: a global network of 3D curves and junctions

Isolated point features + dense multiview stereo

Pro: uncontrolled acquisition; dense texturized models

Con: point cloud; need texture; use a large amount of resources; unscaleable; oversmoothing; lack semantic info

We favor a middle ground approach based on curves

More distinctive features than points, allowing for applications such as 3D modeling and object matching

More efficient in space and time compared to volumetric or mesh-oriented approaches; prioritizes informative areas

More flexible when there aren't enough feature points or texture, on its own or by constraining surfaces

**EXPERIMENTS**

![Experiments images]

**QUANTITATIVE VALIDATION**

![Validation graphs]

**ONGOING WORK: LOFTING**

![Lofting images]

**CONCLUSION**

- Global multiview reconstruction based on image curve content
- Resolution anchored at singularities for progressive & crisp reconstructions

Multiview-3d-Drawing.sf.net
Shape from Selfies: Human Body Shape Estimation using CCA Regression Forests

Endri Dibra, Cengiz Öztireli, Remo Ziegler, Markus Gross

Contributions

Goal: Fast and Automatic Human Shape Estimator from Silhouettes
- Require no pose estimation or known camera calibration
- Assume mild self occlusion (e.g., Selfie like poses)
- Automatic mapping from silhouette to 3D body shape in ms

Proposed Solution
- Extract novel silhouette features capturing global and local information, robust to silhouette noise
- Classify Viewing Direction with a Random Forest Classifier
- Project features from two complementary views into correlated CCA spaces to obtain more discriminative ones
- Train a Random Forest Regressor from projected features to 3D mesh spaces learned from Scape

Data Generation

Scape [1] Generated Meshes
Silhouette Features
Body Measurements
Running & Walking Poses (LBS)

Canonical Correlation Analysis (CCA)

\[
\begin{align*}
\text{arg max} & \, \text{corr}(b_1'X_1, b_2'X_2) \\
& b_1, b_2 \in \mathbb{R}^d \\
\end{align*}
\]
- Noise is removed
- Correlation is maximized
- Lower dimensionality
- Unaffected by rot., trans. and scale of the features

Shape Estimations

Full Pipeline

1D Features
Offline Training
Output Mesh
View Classifier
CCA
Random Forest
Online Testing

Quantitative Comparisons

Table 1: Comparisons to state-of-the-art methods, variations of our method (RF, CCA-RF-1, CCA-RF-2) and ground truth, via various measures illustrated on the right. Errors are represented as Mean±Std. Dev over 800 unseen meshes and are expressed in millimeters. Note that we operate under significantly more general setting than the state-of-the-art methods.

<table>
<thead>
<tr>
<th></th>
<th>[2]</th>
<th>[3]</th>
<th>[4]</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Head circumference</td>
<td>102±2</td>
<td>102±2</td>
<td>102±2</td>
</tr>
<tr>
<td>B: Neck circumference</td>
<td>118±3</td>
<td>118±3</td>
<td>118±3</td>
</tr>
<tr>
<td>C: Shoulder-blade/sling length</td>
<td>43±3</td>
<td>43±3</td>
<td>43±3</td>
</tr>
<tr>
<td>D: Chest circumference</td>
<td>106±2</td>
<td>106±2</td>
<td>106±2</td>
</tr>
<tr>
<td>E: Waist circumference</td>
<td>223±3</td>
<td>223±3</td>
<td>223±3</td>
</tr>
<tr>
<td>F: Pelvis circumference</td>
<td>115±2</td>
<td>115±2</td>
<td>115±2</td>
</tr>
<tr>
<td>G: Wrist circumference</td>
<td>91±2</td>
<td>91±2</td>
<td>91±2</td>
</tr>
<tr>
<td>H: Hip circumference</td>
<td>173±2</td>
<td>173±2</td>
<td>173±2</td>
</tr>
<tr>
<td>I: Forearm circumference</td>
<td>168±2</td>
<td>168±2</td>
<td>168±2</td>
</tr>
<tr>
<td>J: Arm length</td>
<td>156±2</td>
<td>156±2</td>
<td>156±2</td>
</tr>
<tr>
<td>K: Ins. leg length</td>
<td>62±7</td>
<td>62±7</td>
<td>62±7</td>
</tr>
<tr>
<td>L: Thigh circumference</td>
<td>91±2</td>
<td>91±2</td>
<td>91±2</td>
</tr>
<tr>
<td>M: Calf circumference</td>
<td>97±2</td>
<td>97±2</td>
<td>97±2</td>
</tr>
<tr>
<td>N: Ankle circumference</td>
<td>98±3</td>
<td>98±3</td>
<td>98±3</td>
</tr>
<tr>
<td>O: Overall height</td>
<td>93±2</td>
<td>93±2</td>
<td>93±2</td>
</tr>
<tr>
<td>P: Shoulder breadth</td>
<td>92±7</td>
<td>92±7</td>
<td>92±7</td>
</tr>
</tbody>
</table>

References

Can We Jointly Register and Reconstruct Creased Surfaces by Shape-from-Template Accurately?

Mathias Gallardo, Toby Collins and Adrien Bartoli
ALCoV-ISIT, UMR 6284 CNRS / Université d’Auvergne, Clermont-Ferrand, France

Context and Motivations

Shape-from-Template:
3D reconstruction using the apparent motion of features between two single images and a 3D textured template

Needs textured surfaces

Assumes smooth deformations

• Usual regularizers [2,3]

• Reduction dimensionality [4]

We tackle the problem of reconstructing surface creases

Problem Modeling

Deformation parameters

\[ x = (x_1, \ldots, x_N) \]

• Embedded by \( \mathcal{F} \)

• Triangular regular density mesh

allows complex deformation modeling

Bending Constraint \( C_{\text{bend}} \)

• M-estimator allows piecewise smooth 3D reconstructions by reducing influence of high errors

\[ C_{\text{bend}}(x) = \int_{\Omega} \rho \left( \nabla^2 y(u(x)) \right) \, d\Omega, \]

In red, high error in bending constraint

Comparison of \((\ell_1, \ell_2)\) and Haber M-estimators (bending weight and Haber hyper-parameter)

M-estimator retained: \((\ell_1, \ell_2)\)

\[ \rho(y) = 2 \left( 1 + |y|^{\ell_2} + 1 \right) \]

Boundary Constraint \( C_{\text{bound}} \)

• Projects the 3D surface boundaries in boundaryness map → potential well

\[ C_{\text{bound}}(x) = \int_{\partial \Omega} \rho \left( \nabla^2 y(u(x)) \right) \, d\Omega, \quad \text{with} \quad \rho = \exp \left( -|y|^\ell \right) \]

Enhancement of boundaryness map: color-based foreground detector to reduce false boundary edges (from background clutter or texture)

Solution Strategy

• Initialization with an existing solution [3]

• Non-convex refinement

• Minimization of cost function \( C(x) \) for a dense mesh of \( \mathcal{F}[10^4] \) vertices

• Gauss-Newton optimization with sparse Cholesky solver for normal equations

• Two-stage optimization with images pyramid for boundary constraint

Conclusion

• Modeling and optimization framework to register and reconstruct accurately smooth and creased 3D surfaces from a single image and a deformable 3D template

• Creases modeled by a dense mesh with a robust bending constraint led by an M-estimator

• Use of boundary constraint for a more accurate registration and color-based foreground detector to improve convergence

• Future works: arbitrary topologies and dynamic crease modeling

Previous Attempts

Closest work: convex formulation of [1]

• Correspondences are not sufficiently informative

• Does not use smoothing

Input image

Reconstruction from [1]

Contributions

(i) implicitly model surface creases without knowing a priori their location through an M-estimator for bending constraint

(ii) introduce boundary constraint to complement motion constraint which is sparse

(iii) use statistical color models to help disambiguate non-boundary edges

Global Cost Function

\[ C(x) = \]

\[ C_{\text{crp}}(x) \quad \text{Reduces reprojection error of feature-correspondences} \]

\[ + \lambda_{\text{bound}} C_{\text{bound}}(x) \quad \text{Aligns } \Omega_B \text{ to wherever it is visible in the input image} \]

\[ + \lambda_{\text{iso}} C_{\text{iso}}(x) \quad \text{Prevents the surface from extension and contraction (isometry)} \]

\[ + \lambda_{\text{bend}} C_{\text{bend}}(x) \quad \text{Penalizes non-smooth surfaces and reduces the energy at creases} \]

M-estimators

• norms that fit to outliers/reduce the impact of high errors

\[ C_{\text{crp}} \text{ to handle mis-matches} \]

\[ C_{\text{bound}} \text{ to handle little contrast at edges surface} \]

\[ C_{\text{bend}} \text{ to allow piecewise smooth reconstructions} \]

\[ \min \sum_i \phi(\text{residual}_i) \]

Results

• Real datasets with high-precision (<1 mm) ground truth with structured-light system

• Comparison to SfS state-of-the-art: [1,2,3,4]

• Smooth datasets presented in [3]: small improvement of the 3D mean error

• Quantitative evaluations all over the surface and at the creases: 3D errors and normals

• Better 3D reconstruction at creased regions and all over the surface

References

[1] Saltmarche and Fox, PAMI 2011

[2] Bandy et al., PAMI 2011

[3] Chokheli et al., CVPR 2014

Distractor-supported single target tracking in extremely cluttered scenes

Jingdong Xiao\(^1\) (shine636363@ sina.com), Lintao Gao\(^2\), Rustam Stolkin\(^3\), Alex Leonardis\(^4\)
\(^1\)University of Birmingham, Birmingham, UK
\(^2\)National University of Defense Technology, China

Abstract

We present a novel method for single target tracking in RGB images under conditions of extreme clutter and camouflage, including frequent occlusions by objects with similar appearance as the target.

1. Introduction

Research problem: How to track a target which moves through scenes featuring several other very similar objects?

We propose:

• Multi-level clustering-based robust estimation for online detection and learning of multiple target-like regions (distractors) when they appear near to the true target.
• To distinguish the target from these distractors, we exploit a global dynamic constraint (derived from the target and the distractors) in a feedback loop to improve single target tracking performance when the target is camouflaged in highly cluttered scenes.

2. Our method

The proposed algorithm:

1) simultaneously tracks the target and the distractors using the proposed robust estimation method;
2) extracts a global dynamic model from the relative target and distractor trajectories;
3) feeds the global dynamic information back to the single target tracker to help identify the true target and infer occlusion situations.

3. Results

Evaluation on highly cluttered scenes

Datasets. We have selected 28 highly cluttered sequences from publicly available datasets [1-3].

A. Evaluation of the tracker sub-components

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Overall</th>
<th>Colour</th>
<th>Shape</th>
<th>Camouflage motion</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>6.264</td>
<td>5.6077</td>
<td>6.6777</td>
<td>5.7518</td>
<td>6.1285</td>
</tr>
<tr>
<td>R+H</td>
<td>5.8080</td>
<td>5.2776</td>
<td>7.2067</td>
<td>6.4845</td>
<td>7.2580</td>
</tr>
<tr>
<td>R+H+GDC</td>
<td>7.3545</td>
<td>6.7536</td>
<td>7.2065</td>
<td>6.5659</td>
<td>7.2558</td>
</tr>
</tbody>
</table>

B. Overall performance comparison

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Overall</th>
<th>Colour</th>
<th>Shape</th>
<th>Camouflage motion</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CF</td>
<td>1.251</td>
<td>2.2956</td>
<td>2.7984</td>
<td>1.9619</td>
<td>2.5481</td>
</tr>
<tr>
<td>EF</td>
<td>2.3745</td>
<td>4.2538</td>
<td>6.5910</td>
<td>4.0052</td>
<td>4.9520</td>
</tr>
<tr>
<td>Track</td>
<td>2.3847</td>
<td>2.9844</td>
<td>13.3576</td>
<td>13.9341</td>
<td>15.4405</td>
</tr>
<tr>
<td>HCF</td>
<td>7.0762</td>
<td>7.0315</td>
<td>7.2508</td>
<td>7.5468</td>
<td>7.3688</td>
</tr>
<tr>
<td>Ours</td>
<td>7.9880</td>
<td>8.1886</td>
<td>8.5391</td>
<td>8.0662</td>
<td>8.6388</td>
</tr>
</tbody>
</table>

AUC for the decomposed single target tracking algorithm tested in extremely cluttered scenes. B: baseline algorithm (only colour feature); H: HOG feature; GDC: global dynamic constraint. (Red: best performance; Blue: second best performance).

3. Results (continued)

Evaluation on non-cluttered scenes

We also tested our algorithm on non-cluttered sequences (94 sequences) from OTB100 [1], excluding the already used highly cluttered sequences.

<table>
<thead>
<tr>
<th>Method</th>
<th>Ours</th>
<th>HCF</th>
<th>KCF</th>
<th>Struck</th>
<th>SCm</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>9.78</td>
<td>11.04</td>
<td>9.44</td>
<td>9.34</td>
<td>9.00</td>
</tr>
</tbody>
</table>

4. Conclusions

The proposed algorithm incorporates a novel multi-level clustering method for online detection and learning of target-like contextual image regions (distractors). To disambiguate the target’s path among the distractors, a global dynamic constraint is proposed in a feedback loop to improve the single target tracker. The effectiveness of the algorithm is demonstrated in the experiments.

References


Code available: https://github.com/shine636363/DSTcode

Acknowledgements

This work was funded by EU H2020 RoMaNS 645582 and EPSRC EP/M026477/1 projects. We also acknowledge MoD/Dstl and EPSRC for providing a Department of Defence funded MURI project (EP/N019415/1) which supported the involvement of Alex Leonardis.
1. Weakly Supervised Action Labeling

- Challenge: no temporal localization of the actions during training (only order)
- Contributions: (1) first introduce ECTC to efficiently evaluate all alignments
  (2) comparable to fully supervised methods with 1% label

2. Connectionist Temporal Modeling

- Challenge: large number of possible alignments
- CTC [2]: evaluate all possible alignments using DP

3. Extended Connectionist Temporal Classification (ECTC)

- Challenge: much larger space of possible alignments with degenerated alignments
- ECTC: explicitly enforce consistency with the frame-to-frame visual similarities

4. Frame-level Semi-supervised Learning

- Could be extracted from movie scripts or by labeling actions for a small number of frames
- Significantly reduce the alignment space and boosts the performance of our approach

5. Evaluating Complex Activity Segmentation

- The Breakfast Actions Dataset [3]
  - GndTruth
  - Frames
  - Fully supervised
  - ECTC
  - CTC
  - Uniform
  - OCDC
  - OCDC: Ordering Constrained Discriminative Clustering [4]
  - Uniform: uniformly distributing the occurring actions among frames
  - HTK: Hidden Markov Model Toolkit [3]

6. Evaluating Action Detection

- Clips from subset of the Hollywood2 Dataset [4]
- Weakly supervised action detection results

---

Deep Joint Image Filtering
Yijun Li¹, Jia-Bin Huang², Narendra Ahuja², Ming-Hsuan Yang¹
¹University of California, Merced   ²University of Illinois, Urbana-Champaign

Joint image filtering

- Depth upsampling
- Noise Reduction
- Inverse Halftoning
- Texture Removal

Questions:
1. A learning-based approach for joint image filters.
3. A generic filter to handle image data in a variety of domains.

Limitations of existing methods

1. Fail to consider mutual structures.
2. Hand-crafted objective functions.
3. Inefficiency for optimization-based methods.

Contributions

1. Learning to selectively transfer mutual structures from guidance to target.
3. A generic filter to handle image data in a variety of domains.

Network architecture

- CNN_T
- CNN_G

Contributions:

1. A learning-based approach for joint image filters.
3. A generic filter to handle image data in a variety of domains.

Design rationale

- Directly stack the RGB guidance image and target depth image, and feed them through a generic network → poor performance.
- Replace the RGB guidance image with its edge map [Dollár et al., 2013] → good performance.
- End-to-End: Extract structural features from both the target and guidance image. Then, combine them and reconstruct.

Experiments

1. Depth map upsampling

2. Chromaticity map upsampling

3. Saliency map upsampling

4. Inverse halftoning

5. Cross-modal noise reduction

What has the network learned?

- The learned guidance appears like an edge map.
- CNN_T and CNN_G show strong responses to edges from the target and guidance image respectively.
- CNN_G re-organizes the extracted structural features and suppresses inconsistent details.

Experiments:

1. Depth map upsampling

2. Chromaticity map upsampling

3. Saliency map upsampling

4. Inverse halftoning

5. Cross-modal noise reduction

RMSE comparisons for depth map upsampling.

<table>
<thead>
<tr>
<th>Method</th>
<th>Middlebury (#30, [0,255])</th>
<th>NYU v2 (#449, cm)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4x</td>
<td>8x</td>
</tr>
<tr>
<td>He et al., 2010</td>
<td>4.01</td>
<td>7.22</td>
</tr>
<tr>
<td>Fersl et al., 2013</td>
<td>3.39</td>
<td>5.41</td>
</tr>
<tr>
<td>Ham et al., 2015</td>
<td>3.14</td>
<td>5.03</td>
</tr>
<tr>
<td>Ours</td>
<td>2.14</td>
<td>3.77</td>
</tr>
</tbody>
</table>

Training on a RGB/D dataset NYU v2 [Silberman et al., 2012]

- Learning to selectively transfer mutual structures from guidance to target.
- Trained on a RGB/D dataset NYU v2 [Silberman et al., 2012].

- The learned guidance appears like an edge map.
- CNN_T and CNN_G show strong responses to edges from the target and guidance image respectively.
- CNN_G re-organizes the extracted structural features and suppresses inconsistent details.
Efficient Multi-Frequency Phase Unwrapping using Kernel Density Estimation

Felix Järemo Lawin, Per-Erik Forssén, Hannes Ovrén
(felix.jaremo-lawin, per-erik.forssen, hannes.ovren)@liu.se

Introduction

We introduce an efficient method to unwrap multi-frequency phase estimates for time-of-flight ranging. The algorithm generates multiple depth hypotheses and uses a spatial kernel density estimate (KDE) to rank them. We apply the method on the Kinect v2.

- The Kinect v2 is designed for scenes with less than 8m range, but with our method the effective range can be extended.
- When extending the depth range to the maximal value of 18.75m, we get about 52% more valid measurements than existing drivers in libfreenect2 and Microsoft Kinect SDK.
- Runs in ≈ 190Hz on a Nvidia GeForce GTX 760 GPU. Code is available at: http://www.cvl.isy.liu.se/research/datasets/kinect2-dataset/

Method

In amplitude modulated time-of-flight ranging the depth is obtained by measuring the phase shift \( \phi_m \in [0, 2\pi) \) of the modulated signal given the frequencies \( f_m \).

\[
\text{Figure 1: Left: The Kinect v2 sensor, right: wrapped phases for Kinect v2, in the range 0 to 25 meters. Top to bottom: } \phi_0, \phi_1, \phi_2. \text{ The dashed line at 18.75 meters indicates the common wrap-around point for all three phases. Just before this line we have } n_0 = 9, n_1 = 1, \text{ and } n_2 = 14.
\]

- The corresponding unwrapped phase measurements are \( \phi_m + 2\pi n_m \).
- Phase unwrapping means finding the unwrapping coefficients \( n = (n_0, \ldots, n_{M-1}) \).
- Each vector \( n \) corresponds to a hypothesis of the depth \( t \).
- Our method considers several hypotheses for each pixel location and selects the one with the highest kernel density value in a spatial neighbourhood \( N(x) \):

\[
p(t'(x)) = \frac{\sum_{i \in I} w_{ik} K(t'(x) - t'(x_i))}{\sum_{i \in I} w_{ik}},
\]

\[
w_{ik} = g(x - x_k, \sigma)p(t'(x)|n_i)p(t'(a_i)) \text{, } K(x) = e^{-\frac{x^2}{2\sigma^2}}.
\]

The three factors in \( w_{ik} \) are:
- the spatial weight \( g(x - x_k, \sigma) \).
- the unwrapping likelihood \( p(t'(x)|n_i(x)) \).
- the phase likelihood \( p(t'(x)|a_i(x)) \), where \( a_i = (a_0, \ldots, a_{M-1}) \), are the amplitudes.

The final hypothesis selection is then made as:

\[
t^* = \arg \max_{t' \in T} p(t').
\]

Inliers are classified by \( p(t') > T \).

Results

We apply the method to depth decoding for the Kinect v2 sensor, and compare it to the Microsoft Kinect SDK and to the open source driver libfreenect2.

- Ground truth is constructed by fusing many depth frames from 9 different camera poses into one very accurate depth image.
- Raw measurements from the ground truth pose are decoded into depth images using Microsoft, libfreenect2 and our method.
- A point is counted as an inlier when a method outputs a depth estimate that correspond to a correct unwrapping, which we set to be within 30cm from the ground truth.
- We evaluate on the kitchen dataset with maximal depth of 6.7m and the lecture dataset with maximal depth of 14.6m.

\[
\text{Figure 2: Inlier and outlier rate plots and corresponding ground truth depth images. Each point or curve is the average over 25 frames. In the kitchen (depth limited) curve the algorithms assume a maximum depth of 8m, which simplifies the outlier rejection.}
\]

- A clear improvement can be observed in the depth images, especially in large depth scenes.

\[
\text{Figure 3: Single frame output. Left: libfreenect2, Center: proposed method. Right: corresponding RGB image. Pixels suppressed by outlier rejection are shown in green.}
\]

- Recordings of raw Kinect v2 measurements were unwrapped and passed to the Kinect fusion implementation KinFu in the Point Cloud Library.

\[
\text{Figure 4: KinFu scans of two different scenes using depth images produced by libfreenect2 and the proposed method. The input duration was 200 frames.}
\]
A Multi-Scale CNN for Affordance Segmentation in RGB Images

Anirban Roy and Sinisa Todorovic

Problem Statement

Input image

Walkable  Sittable  Layable  Reachable  Movable

Pixel-wise affordance prediction

Challenges

- Cluttered indoor scene
- Distorted 3D geometry due to camera projection
- Lack of pixel-wise annotated training data

Contributions

- Reliable affordance prediction using mid-level cues
- Multi-scale CNN to capture high and low level details
- New affordance dataset with pixel-wise labels

Semi-Automated Ground-Truth Generation

Input

RGB image  Depth map  Object labels

Surface normals

Fitted planes

Ground-truth affordance maps

- Compute surface normals from the depth map
- Fit planes using RANSAC
- Affordance from input object labels and planes
- Manually correct affordance labels

Multi-Scale CNN for Affordance Prediction

- We extract mid-level cues from a RGB (no depth)
- Three multi-scale CNNs predict mid-level cues
- The mid-level cues are combined by another multi-scale CNN for affordance prediction

Pixel-wise Jaccard index on NYUv2

<table>
<thead>
<tr>
<th>Without depth</th>
<th>68.23</th>
<th>31.44</th>
<th>36.10</th>
<th>57.24</th>
<th>43.84</th>
<th>46.37</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without surface normals</td>
<td>64.36</td>
<td>32.32</td>
<td>37.77</td>
<td>57.70</td>
<td>44.64</td>
<td>47.36</td>
</tr>
<tr>
<td>Without object labels</td>
<td>62.24</td>
<td>32.42</td>
<td>37.84</td>
<td>58.28</td>
<td>41.70</td>
<td>46.50</td>
</tr>
<tr>
<td>Without mid-level cues</td>
<td>58.45</td>
<td>24.63</td>
<td>31.20</td>
<td>50.54</td>
<td>34.20</td>
<td>39.80</td>
</tr>
<tr>
<td>With ground truth cues</td>
<td>70.43</td>
<td>37.61</td>
<td>43.33</td>
<td>63.14</td>
<td>51.37</td>
<td>53.23</td>
</tr>
<tr>
<td>Our approach</td>
<td>66.74</td>
<td>34.44</td>
<td>40.18</td>
<td>60.01</td>
<td>46.42</td>
<td>49.56</td>
</tr>
</tbody>
</table>

Affordance Prediction on NYUv2

Input image  Walkable  Sittable  Layable  Reachable  Movable

Ground-truth affordance maps

Acknowledgment: This work was supported in part by grant NSF RI 1302700.
Hierarchical Dynamic Parsing and Encoding for Action Recognition

Bing Su¹, Jiahuan Zhou², Xiaqing Ding³, Hao Wang¹, Ying Wu²

¹Institute of Software Chinese Academy of Sciences. ²Northwestern University. ³Tsinghua University.

Motivation

- Representation matters:
  - The performance of action recognition methods depends heavily on the representation of video data.

- Dynamics are inherent:
  - Dynamics characterize the inherent temporal dependencies of actions.
  - Existing methods either cannot directly lead to vector representations with a fixed dimension, or treat the changes of all successive frames equally.

- Dynamics are not uniform:
  - The dynamic behind an action is time-varying, non-stationary and has some intuitive rhythms or regularities.
  - Humans can recognize an action from some ordered key frames or poses. These key poses segment the whole action into different divisions, and each division consists of the frames related to a key pose.

The dynamics of an action can be also viewed as a hierarchy. The dynamics within each stage are relatively stable, and the dynamics of the sequence of the stages represent the essential evolution of the action.

Dynamic Parsing

- Learn the parse of an action sequence from the sequence itself:
  - Capture the temporal structures w.r.t. relatively-uniform local dynamics.

- Unsupervised Temporal Clustering:
  - Represent the video with a sequence of frame-wide features
  - A partition of $X$ is defined by a segmentation path $P = [p_1, p_2, \cdots, p_L]$. $p_l = [s_l, e_l]^T$ denotes the start and end frames of the $l$-th division; $f$ controls the maximum extent of wrapping
  - Define the essential sequence of $X$: $U = \{\mu_1, \mu_2, \cdots, \mu_L\}$ which can be viewed as the sequence of key poses.

Objective:

- Minimize $F(U)$

Optimization:

- Given $P$, optimize $U$: compute the means of all the divisions

Hierarchical Dynamic Encoding

- Incorporate the dynamics in the hierarchy of two layers into a joint representation

The first layer modeling:

- The action sequence is parsed into several smooth-changing divisions corresponding to different key poses or temporal structures by unsupervised temporal clustering
- The dynamics within each stage are encoded by mean-pooling (M-HDPE) or rank pooling (R-HDPE).

The second layer modeling:

- The dynamics of the ordered representations extracted from the previous layer is encoded again by rank pooling to form the overall representation

Contribution

- The proposed HDPE is a new unsupervised representation learning method. It hierarchically abstracts the prominent dynamic and generates a representation that is robust to speed and local variations.
- We propose an unsupervised method for temporal clustering to achieve efficient dynamic parsing.
Distinct Class-specific Saliency Maps for Weakly-supervised Semantic Segmentation

Wataru Shimoda  Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

Objective

Weakly supervised segmentation
- Use only image-level annotation

Fully supervised annotation

Person
horse
Car

Subtraction of class-specific derivatives

For multi-class images
- Only small differences were observed among the derivatives of the different classes

Assumption
- Raw saliency maps are affected by both class-specific saliency and generic object-ness
- The degree of class saliency factors should be larger than the generic object-ness factor.
- Background regions do not respond.

Subtraction
- Subtract the derivatives among the different classes.
- Interestingly, in most of the cases, we obtained much cleaner class maps than raw maps.
- The improved class saliency maps $M^c_{xy}$ with respect to class $c$ is computed by:

$$
M^c_{xy} = \sum_{c'} (M_{xy}^c - M_{xy}^{c'}) \quad \text{for} \quad c \neq c'.
$$

Weakly supervised segmentation
- Improved the method by Simonyan et al. [1] greatly
- Achieved state-of-the-art in weakly-supervised segmentation with PASCAL VOC 2012

Contributions
- Improved the method by Simonyan et al. [1] greatly
- Achieved state-of-the-art in weakly-supervised segmentation with PASCAL VOC 2012

BP-based Visualization

Visualize class-specific saliency maps based on the derivatives of the class scores with respect to the input image
- proposed by K. Simonyan et al. at ICLR 2014 [1]
- Visualize contributed pixels on CNN classification
- Use derivatives obtained by back-propagation

CNN Architecture

Improved points
- (each point contributes 1-3pt improvement)
- Fully Convolutional Net
- Guided back propagation [2]
- Use the derivatives of multiple intermediate layers
- Aggregate multi-scale class saliency maps (3 scales)

We back-propagate expected class scores generated by setting 1 for one of the top-N classes and 0 for the others. $w^c_{i,x}$ represents up-sampled i-th layer derivative which is obtained by propagating class scores from the top layer.

Each class saliency maps $M^c_{i,x,y}$ is calculated by:

$$
M^c_{i,x,y} = \max_{h_{i}} \left| h_{i}(x,y,k) \right|
$$

where $h_{i}(x,y,k)$ is the index of the element of $w^c_{i}$.

- Fine-tune full-conv VGG-16 network with Sigmoid cross entropy loss with random-resized images (300-700px)
- Sum up the derivatives of Conv3, Conv4, and Conv5
- Aggregate the class maps of 400*400, 500*500, and 600*600.

Experiments

Dataset: PASCAL VOC 2012 + trainaug [3]

Comparison with Simonyan et al. [1]
- Our gradient maps visualize class regions clearly.
- We applied FC-CRF to saliency maps obtained by Simonyan et al.[1] in the same way.
- The margin was more than 10%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sim et al. + CRF</td>
<td>33.8</td>
</tr>
<tr>
<td>Ours</td>
<td>44.2</td>
</tr>
</tbody>
</table>

Comparison with state-of-the-arts
- A means using additional images.
- B means using additional supervision.

<table>
<thead>
<tr>
<th>Method</th>
<th>A</th>
<th>B</th>
<th>Mean IOU</th>
</tr>
</thead>
<tbody>
<tr>
<td>One point (ECCV2016)</td>
<td>-</td>
<td>✓</td>
<td>46.1</td>
</tr>
<tr>
<td>Check Mask (ECCV2016)</td>
<td>-</td>
<td>✓</td>
<td>51.5</td>
</tr>
<tr>
<td>MIL-FCN (ECCV2015)</td>
<td>-</td>
<td>-</td>
<td>25.7</td>
</tr>
<tr>
<td>EM-Adapt (ECCV2015)</td>
<td>-</td>
<td>-</td>
<td>38.2</td>
</tr>
<tr>
<td>CCNN (ECCV2015)</td>
<td>-</td>
<td>-</td>
<td>34.5</td>
</tr>
<tr>
<td>MIL-seg (CVPR2015)</td>
<td>✓</td>
<td>-</td>
<td>42.0</td>
</tr>
<tr>
<td>STC (arXiv:1509.03150)</td>
<td>✓</td>
<td>-</td>
<td>49.8</td>
</tr>
<tr>
<td>SEC (ECCV2016)</td>
<td>-</td>
<td>-</td>
<td>50.7</td>
</tr>
<tr>
<td>Ours w/o CRF</td>
<td>-</td>
<td>-</td>
<td>40.5</td>
</tr>
<tr>
<td>Ours w/ CRF</td>
<td>-</td>
<td>-</td>
<td>44.2</td>
</tr>
</tbody>
</table>

References


Source code

https://github.com/shimoda-uec/dcsdm

Project page

http://mm.cs.uuec.ac.jp/shimoda-k-space/dcsdm
Automatic Attribute Discovery with Neural Activations

Sirion Vittayakorn\textsuperscript{1}, Takayuki Umeda\textsuperscript{2}, Kazuhiko Murasaki\textsuperscript{2}, Kyoko Sudo\textsuperscript{2}, Takayuki Okatan\textsuperscript{3}, Kota Yamaguchi\textsuperscript{3}
\textsuperscript{1}UNC at Chapel Hill, NC, USA \textsuperscript{2}NTT Media Intelligence Laboratories, Japan \textsuperscript{3}Tohoku University, Japan
http://cs.unc.edu/~sirionv

Overview

- Can we learn visual attributes without a supervised dataset?
- Our approach: look at neurons
- KL divergence helps identify visual attributes from noisy label association
- Relation to human perception
- Neurons to saliency detection

KL divergence identifies prime units

![Diagram showing KL divergence](image)

- Some neurons in the pre-trained network respond to visual attributes, even if they are not explicitly supervised
- KL-divergence on activation histogram can identify such neurons (prime units)

Prime-units perceive attributes like humans do

- Learning a classifier on top of prime units shows close proximity to humans
- Evaluating visualness to human-agreement correlation

Visualness \(\mu([\text{classifier}] D_\alpha^p, D_{\text{human}})\) = accuracy(classifier, \(D_{\text{human}}\))

Layers characterize attribute perception

- From early primitive attributes to deeper abstract attributes
- Human agreement highest in the middle?
- Fine-tuning affects maximum KL ratio per layer

Automatically discovered attributes

![Diagram showing attribute discovery](image)

- KL-dominant neurons correspond to salient region with a human-agreement correlation

Prime units can identify salient region

- Accumulating sliding mask [Zhou 14] by largest KL units
- Baseline: smoothed gradients [Simonyan 2014]

Pixel-wise performance evaluation against human annotation
**Problem**

Novel view synthesis: given an input image, synthesizing new images of the same object or scene seen from novel viewpoints.

**Why Appearance Flow**

1) Avoids perceptual blurriness caused by naive $L_p$ loss minimization – no longer allowed to predict the ‘mean’ that minimizes the error but loses high-frequency details.

2) Color identity is preserved by construal – can only use existing pixels.

3) Interpretable results – can visualize exactly how each output image is constructed.

**Method**

Key insight: high visual correlation across different views $\rightarrow$ pixels to be synthesized likely exist in the input view.

Appearance flow: 2-D coordinate vectors specifying where to copy pixels to reconstruct the target view.

Learning to predict appearance flow:

$$\min_{i \in C \times T} \sum_{i \in C \times T} ||I(i, T) - s(i, T)||_p$$

The above constraint can be rewritten in the form of bilinear sampling [2]:

$$\hat{s}(i, T) = \sum_{p \in \text{neighbors of } (i, s(i, T))} P(i, p) (1 - \hat{x}(i, p)) (1 - \hat{y}(i, p))$$

Learning to leverage multiple input views:

1. Extra output channel for confidence of each single-view prediction.
2. Final prediction = per-pixel weighted sum (weight determined by confidence) over all single-view predictions.

**ShapeNet Results**

Data setup:

- 7,497 cars and 700 chairs split into 80% training and 20% testing.
- Each shape is rendered for 504 viewing angles (azimuth $= 0^\circ$ -- $355^\circ$, elevation $= 0^\circ$ -- $30^\circ$ both at steps of $5^\circ$)
- Viewpoint transformation limited to azimuth only with the range of $-180^\circ$ -- $180^\circ$ at steps of $20^\circ$.

Quantitative measure (mean $L_1$ pixel error):

<table>
<thead>
<tr>
<th>Method</th>
<th>Car Error</th>
<th>Chair Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-view</td>
<td>0.560</td>
<td>0.492</td>
</tr>
<tr>
<td>Tatarchenko et al. [1]</td>
<td>0.404</td>
<td>0.345</td>
</tr>
<tr>
<td>Ours</td>
<td>0.385</td>
<td>0.344</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Car Error</th>
<th>Chair Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-view</td>
<td>0.248</td>
<td>0.409</td>
</tr>
<tr>
<td>Tatarchenko et al. [1]</td>
<td>0.285</td>
<td>0.421</td>
</tr>
<tr>
<td>Ours</td>
<td>0.285</td>
<td>0.421</td>
</tr>
</tbody>
</table>

**Kitti Results**

Data setup:

- 11 driving sequences through urban scenes (9 for training, 2 for testing)
- Viewpoint transf. = car ego-motion
Top-down Learning for Structured Labeling with Convolutional Pseudoprior

Saining Xie* and Xun Huang* and Zhuowen Tu
(sxie.ztu)@ucsd.edu, xh258@cornell.edu
University of California San Diego, Cornell University

Abstract

Current practice in convolutional neural networks (CNN) remains largely bottom-up and the role of top-down process in CNN for pattern analysis and visual inference is not very clear. In this paper, we propose a new method for structured labeling by developing convolutional pseudoprior (ConvPP) on the ground-truth labels. Our method has several interesting properties: (1) compared with classic machine learning algorithms such as CRFs and Structural SVM, ConvPP automatically learns rich convolutional kernels to capture both short- and long-range contexts; (2) compared with cascade classifiers like Auto-Context, ConvPP avoids the iterative steps of learning a series of discriminative classifiers and automatically learns contextual configurations; (3) compared with recent efforts combining CNN models with CRFs and RNNs, ConvPP learns convolution in the labeling space with improved modeling capability and less manual specification; (4) compared with Bayesian models like MRFs, ConvPP capitalizes on the rich representation power of convolution by automatically learning priors built on convolutional filters. We accomplish our task using pseudo-likelihood approximation to the prior under a novel fixed-point network structure that facilitates an end-to-end learning process.

Background

• We approximate the appearance \( p(X|Y) \) with a discriminative model
• And the prior \( p(Y) \) with pseudo-prior
  • Detailed approximation steps please refer to our paper

\[
p(Y|X) \propto p(Y)p(X|Y) \prod_i p(y_i|y_{N_i}) \prod_i p(y_i|X)
\]

Bottom-up CNN

Top-down convolutional pseudo-prior (ConvPP)

Method

Illustration of ConvPP learning process in 1-D structured labeling tasks (OCR)

Integrating with Fully Convolutional Networks (FCN) for 2-D structured labeling tasks

Network Architecture

Training Stage 1

Training Stage 2

Training Stage 3

Testing

Results

• Structured Labeling: 1-D case

• Image Labeling: 2-D case

Performance on OCR dataset:

Performance on OCR dataset:
Variing training size:

Performance on OCR dataset:
[1] Varying context window length
[2] Varying testing iterations

Performance on SIFT Flow dataset

Iterative update of labeling results during testing. Segmentation results are gradually refined.
Generative Image Modeling using Style and Structure Adversarial Networks
Xiaolong Wang, Abhinav Gupta
Carnegie Mellon University

**Generative Adversarial Network (GAN)**

- Higher Quality of Images
- More Interpretable

**Image Formation: Structure + Style**

- Output1: Surface Normal
- Output2: Natural Scenes

**The Generator**

- The Generator
- Learning with FCN Constraints

**Style-GAN**

- Style Generator Network
- Fully Convolutional Network
- Surface Normal Estimation
- Binary Classification
- Real Images

**Joint Learning**

- Structure Discriminator Network
- Style Discriminator Network
- Generated Normals

**3D Rendering Engine**

- Input
- Output
- Original

**Generated Normals and Images**

- Uniform Noise Distribution

**Walking in the Latent Space**

- Fix Style & Change Structure
- Fix Style & Change Style

**Quantitative Evaluation**

- Apply ImageNet pre-trained detector/classifier on generated images
- MAP
- Maximum Norm

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>S^2-GAN</th>
<th>DCGAN</th>
<th>GIST</th>
<th>Places-AlexNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>30.9</td>
<td>-</td>
<td>21.3</td>
<td>19.7</td>
<td>38.1</td>
</tr>
<tr>
<td>S^2-GAN</td>
<td>32.4</td>
<td>19.1</td>
<td>20.1</td>
<td>27.7</td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>37.6</td>
<td>35.3</td>
<td>27.1</td>
<td>23.0</td>
<td>39.0</td>
</tr>
</tbody>
</table>

(a) Detection on NYUv2 Dataset
(b) Classification on SUN RGB-D dataset.

Code Available!
Joint Learning of Semantic and Latent Attributes

Peixi Peng¹, Yonghong Tian¹, Tao Xiang², Yaowei Wang³, Tiejun Huang¹
¹National Engineering Laboratory for Video Technology, Peking University, China
²School of EECS, Queen Mary University of London, UK
³Department of Electronic Engineering, Beijing Institute of Technology, China
{pxpeng, yhtian, tjhuang}@pku.edu.cn, yaoweiwang@bit.edu.cn, t.xiang@qmul.ac.uk

1 Motivations

- Attributes are effective mid-level representations for cross-class transfer learning problems such as zero-shot learning (ZSL) and person re-identification (Re-ID).
- Joint learning of semantic and latent attributes (both discriminative and non-discriminative/background ones) are important:
  - More comprehensive discriminative representations
  - Better attributes prediction with background attributes

Contributions

- A unified framework for learning both user-defined semantic attributes and discriminative latent attributes is proposed.
- We further develop a novel dictionary learning model which decomposes the learned dictionary subspace into three parts corresponding to the semantic, discriminative latent as well as background latent attributes respectively.

2 Methodology

Decomposition of the dictionary:
- UDAC (D¹): correlated to the semantic attributes
- D-LA (D²): the discriminative latent attributes correlated to the class labels
- B-LA (D³): the background latent attributes which capture all the residual information

The stepwise form:
- The minimization of the first reconstruction error term enables UDAC and D-LA to encode the feature as much as possible.
- The minimization of the second reconstruction error term enables B-LA to encode and align the residual part.
- The combination of these two terms is to prevent the B-LA from dominating the reconstruction error leading to trivial solutions for UDAC and D-LA.

Algorithm 1: The proposed algorithm

Input: X₀, initial dictionary D₀, D¹, D², W. randomly, X₀ ∈ [0, X₀] → Φ; Output: D₀, D¹, D², X₀, X¹, X², W.
while Non-convergence do
  Coding problem: compute code X₀, compute code X¹, compute code X².
  Updating dictionaries: update D₀ and D¹, update D².
  Updating W: update W.

3 Experiments

- Zero-shot Learning and Attribute Prediction (ATT)
- Person Re-ID
- Visualization of the learned D-LA

<table>
<thead>
<tr>
<th>DATASET</th>
<th>AwA</th>
<th>CUB</th>
<th>AwA</th>
<th>CUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAP</td>
<td>57.5</td>
<td>-</td>
<td>72.8</td>
<td>61.8</td>
</tr>
<tr>
<td>ALE</td>
<td>43.5</td>
<td>18.0</td>
<td>65.7</td>
<td>60.3</td>
</tr>
<tr>
<td>UMF</td>
<td>48.6</td>
<td>18.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CSHAP</td>
<td>45.6</td>
<td>17.5</td>
<td>74.3</td>
<td>68.7</td>
</tr>
<tr>
<td>SSE</td>
<td>76.3</td>
<td>30.4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SIE</td>
<td>61.9</td>
<td>40.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>JLSE</td>
<td>80.5</td>
<td>42.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LatEn</td>
<td>76.1</td>
<td>51.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>82.1</td>
<td>56.5</td>
<td>73.6</td>
<td>78.3</td>
</tr>
</tbody>
</table>

The accuracy and mAUC are reported for ZSL and ATT respectively. There is a performance gain compared to the published version of this paper since we optimize the implementation method. For more details see the upcoming open codes.

VIPeR | PRID | Market | iLIDS |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RPLM</td>
<td>27.0</td>
<td>15.0</td>
<td>-</td>
</tr>
<tr>
<td>kCCCA</td>
<td>37.0</td>
<td>15.0</td>
<td>-</td>
</tr>
<tr>
<td>MFA</td>
<td>39.6</td>
<td>20.9</td>
<td>37.6</td>
</tr>
<tr>
<td>kLFDA</td>
<td>39.9</td>
<td>21.6</td>
<td>38.4</td>
</tr>
<tr>
<td>XQDA</td>
<td>40.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MLAPG</td>
<td>40.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours_L</td>
<td>41.3</td>
<td>26.8</td>
<td>47.4</td>
</tr>
<tr>
<td>aMTL</td>
<td>42.3</td>
<td>18.0</td>
<td>-</td>
</tr>
<tr>
<td>Ours_U</td>
<td>28.4</td>
<td>16.3</td>
<td>-</td>
</tr>
<tr>
<td>Ours_All</td>
<td>45.4</td>
<td>26.8</td>
<td>-</td>
</tr>
</tbody>
</table>

The rank 1 is reported. Ours_L means only D-LA are used, Ours_U means only semantic attributes are used, and Ours_All means both type of attributes are used for person Re-ID.

Most are visually meaningful in a subtle way and thus may have been ignored by human annotators, such as clothes with stripes, white logo on the chest, white coat with dark clothes and female with long hairs and short sleeve.
A Unified Multi-scale Deep Convolutional Neural Network for Fast Object Detection

Zhaowei Cai¹, Quanfu Fan², Rogerio Feris³, and Nuno Vasconcelos¹
¹UC San Diego, ²IBM Watson Research

I. Introduction

- Motivations:
  - There is an inconsistency between the sizes of objects, which are variable, and filter receptive fields, which are fixed, in Faster-RCNN framework.
  - Multi-scale detection is not well addressed in CNN based object detection frameworks.
  - The original input images are usually upsampled to boost performance, which exponentially increases the memory and computation costs of the detector.

- Contributions:
  - This work proposes a unified multi-scale deep CNN, denoted the multi-scale CNN (MS-CNN), for fast object detection.
  - To ease the inconsistency between the sizes of objects and receptive fields, object detection is performed with multiple output layers, each focusing on objects within certain scale ranges.
  - Feature upsampling (implemented by a deconvolutional layer) is used as an alternative to input upsampling, which improves detection accuracy but adds trivial computation and no parameter.

II. Multi-scale Object Detection

- Inspired by previous evidence on the benefits of the strategy of (c) over that of (b), we propose a new multi-scale strategy (g). This can be seen as the deep CNN extension of (c), but only uses a single scale of input.

III. Multi-scale Object Proposal Network

- Each detection branch detects objects that match its scale, and the combination of those branches forms a strong multi-scale detector.
- Objective function:
  \[ \mathcal{L}(W) = \sum_{m=1}^{M} \sum_{i \in S^m} \alpha_m l^m(X_i, Y_i | W) \]
  where \( l(X, Y | W) = L_{cls}(p(X), y) + \lambda |y \geq 1| L_{loc}(b, \hat{b}) \)

IV. Object Detection Network

- Unified objective function:
  \[ \mathcal{L}(W, W_d) = \sum_{m=1}^{M} \sum_{i \in S^m} \alpha_m l^m(X_i, Y_i | W) + \sum_{i \in S^m} \alpha_d l^d(X_i, Y_i | W, W_d) \]
  - Trunk CNN layers are shared with proposal sub-network.
  - ROI pooling is applied to the top of the "conv4-3" layer.
  - A deconvolutional layer is used to upsample feature maps as an alternative of input upsampling, avoiding issues such as large memory requirements, slow training and testing.
  - Object and context regions are stacked together immediately after ROI pooling, followed by an extra convolutional layer to compress redundant information and avoid parameters increase.

V. Experimental Results

- Datasets
  - KITTI: 7,481 images (1250×375) for training and 7,518 for testing, no testing ground truth is available.
  - Caltech: 32,077 images (640×480) for training and 4,024 for testing.

- Proposal comparison
  - achieves a recall about 98% with only 100 proposals of high quality.

- Ablation study
  - input size, feature upsampling, context embedding

- Comparison with
  - Faster-RCNN
  - Regionlets
  - 3DOP
  - MS-CNN

- Comparison on KITTI
  - achieves state-of-the-art performance, high detection rate, robust to small and occluded pedestrians.

- Comparison with Caltech
  - achieves state-of-the-art performance, high detection rate, robust to small and occluded pedestrians.

- Real-time running speed
  - up to 10 fps on KITTI (1250×375) and 15 fps on Caltech (640×480) images.

- Reproducible research
  - https://github.com/zhaoweicai/mscnn
Deep Specialized Network for Illuminant Estimation
Wu Shi1, Chen Change Loy1,2, Xiaou Tang1,2
1The Chinese University of Hong Kong
2Shenzhen Institute of Advanced Technology, CAS
{sw015, ccloy, xtang}@ie.cuhk.edu.hk

Introduction
Challenges
• The observed color \( I_c \) is given by
  \[
  I_c = E_c \times R_e, \quad c \in \{r, g, b\}
  \]
  (1)
  where \( E \) is the illumination and \( R \) is the reflectance. Estimating \( E \) from \( I \) is under-determined.
• Searching the large hypothesis space for an accurate illuminant estimation is hard due to the ambiguities of unknown reflections and local patch appearances.

Contributions
• The proposed network combines two networks to work hand-in-hand for robust illuminant estimation.
• A novel notion of ‘branch-level ensemble’ is introduced.
• Through a winner-take-all learning scheme, the two branches of HypNet are encouraged to specialize on estimating illuminants for specific regions.
• SelNet yields much better final predictions than simply averaging the hypotheses.

Methods
The proposed Deep Specialized Network consists of two closely coupled sub-networks:

(1) **HypNet**
• Generates two competing hypotheses for an illuminant estimation of an image patch.
• The ‘winner-take-all’ learning strategy
  \[
  L(\Theta) = \min_{k \in \{A,B\}} (|\bar{E}_k - E^*|^2) \]
  (2)
  where \( \bar{E}_k \) is the estimate of branch \( k \) and \( E^* \) is the ground truth illumination.

(2) **SelNet**
• Generates a score vector to pick the final illuminant estimation from one of the branches in HypNet.
• During training, the ground truth label for SelNet is generated from the ground truth illumination and the two hypotheses obtained from HypNet.

Experimental Results

<table>
<thead>
<tr>
<th>Global-Illuminant Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
<tr>
<td>CCC (5.50)</td>
</tr>
</tbody>
</table>

Table 1: The Color Checker [1, 2] dataset.

<table>
<thead>
<tr>
<th>NUS 8-camera [3] dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
<tr>
<td>CCC (1.40)</td>
</tr>
</tbody>
</table>

Table 2: The NUS 8-camera [3] dataset.

References
Learning Semantic Segmentation with Weakly-Annotated Videos

Goals
- Train a model for semantic segmentation
- Use as little supervision as possible

State-of-the-art
- Based on FCNN framework [Chen et al., ICLR 2015]
- EM-like approaches: E-step - pixel label inference from image tags, M-step - backpropagation
- Label inference requires manual constraints on the size of segments [1, 2] and lacks precision

Overview
- Our approach uses motion as a cue in label inference
- Robust to errors in motion segmentation methods [3]

Our approach
Integrates motion and semantic cues

Loss
Pixel-wise cross entropy loss:
\[ \mathcal{L}(x, p) = - \frac{1}{N} \sum_{i=1}^{N} \log(p_{y_i}(x_i)) \]

Graph-based label inference

Training
Datasets
- YouTubeObjects: 10 categories, 155 videos, 2511 shots
- ImageNet videos: 10 categories, 795 videos, 2120 shots
- Pascal VOC 2012: 20 categories (10 corresponding to YouTubeObjects used on validation set), 1469 images

Training procedure
- Prune shots with erroneous motion segmentations based on segment size
- Train our model on the remaining shots
- Finetune on small subset, selecting one shot per video with previously trained model
- When training with images and videos jointly use [1] for label inference on images

Experimental evaluation
Comparison to [1]
- Our approach shows significant improvements by using motion constraints
- Still images contain additional information

Comparison to the state-of-the-art

Qualitative results

Co-localization
- Replace location unary in [3] with our predictions
- Evaluate co-localization performance on YouTubeObjects

References
1. Motivation

- Existing train-once-and-deploy person re-identification approaches are not scalable, and sometimes even not plausible.
  - Extensive pre-labelled training data
  - Small testing population
- In a real-world, given low Rank-1 recognition rates, human operators are required to verify the unreliable ranking lists.
- How to explore human-in-the-loop for person re-identification?

2. Contribution

- A Human Verification Incremental Learning (HVIL) model which enables human-in-the-loop person re-identification.
  - Less labouring effort
  - Flexible feedback
  - Immediate benefit
- Compared to Post-rank Optimisation (POP) [1], HVIL enables incremental model improvement from cumulative human feedback.
- A Regularised Metric Ensemble Learning (RMEL) model is proposed when human feedback becomes unavailable.

3. Approach

Modelling Human Feedback as a Loss Function

- An incrementally optimised ranking function, \( f_{\alpha}(x^i) : \mathbb{R}^d \rightarrow \mathbb{R}, y \in L = \{m, s, w\} \) as true-match, strong-negative, and weak-negative respectively.

\[
err(f_{\alpha}(x^i), y) = L_\alpha(rank(f_{\alpha}(x^i))),
\]

(1)

- A novel re-ranking loss is introduced:

\[
L_\alpha(k) = \begin{cases} 
\sum_{i=1}^{k-1} \alpha_s & \text{if } y \in \{m, w\}, \\
\sum_{i=k+1}^{n_x} \alpha_s & \text{if } y \in \{s\}, \quad \text{with } \alpha_1 \geq \alpha_2 \geq \cdots \geq 0.
\end{cases}
\]

(2)

- We set \( \alpha_i = \frac{1}{i} \) when \( y \) indicates a true-match, and \( \alpha_i = \frac{1}{n_y-1} \) with \( n_y \) the gallery size when \( y \) represents a weak-negative or strong-negative.

Real-time Model Update for Instant Feedback Reward

- Consider re-ranking model \( f(\cdot) \) as a negative Mahalanobis distance:

\[
f_{\alpha}(x^i) = -\left[ (x^p - x^g)^T M (x^p - x^g) \right], \quad M \in S^d_+.
\]

(3)

- For real-time feedback and reward, \( f(\cdot) \) is estimated on each human feedback in an online manner [2].

\[
M_t = \arg\min_{M \in S^d_+} \Delta_T(M, M_{t-1}) + \eta L^{(t)},
\]

(4)

Metric Ensemble Learning for Automated Re-id

- HVIL learns a series of models \( \{M_j\}_{j=1}^n \) incrementally optimised locally for a set of probes with human feedback. We further propose to learn an ensemble learning model to re-id further probes without human feedback.

\[
f_{\alpha}^{res} = f_{\alpha}^{res}(x^i) = -d_{\alpha}^T W d_{\alpha}, \quad \text{s.t. } W \in S^d_+.
\]

(5)

4. Experiments

Human Feedback Protocol

- Large gallery set: Gallery contains 1000 identities.
- Limited feedback: Maximally 3 rounds of feedback for each probe.
- Limited patience: Users only verify top 5% ranked gallery images.

Human-In-The-Loop Re-ID Performance

- Supervised models were trained by an average of 3,483 cross-view images of 360 identities on CUHK03, and 7,737 images of 501 identities on Market-1501.
- Human-in-the-loop models requires maximally 3 feedbacks per probe.

Table 1: Evaluating human-in-the-loop person re-id with CMC performances.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CUHK03 (( N_m = 1100 ))</th>
<th>Market-1501 (( N_m = 1500 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank (%)</td>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>L2</td>
<td>2.9</td>
<td>31.1</td>
</tr>
<tr>
<td>KLFDA</td>
<td>5.9</td>
<td>47.3</td>
</tr>
<tr>
<td>XQDA</td>
<td>3.7</td>
<td>40.2</td>
</tr>
<tr>
<td>MLAPG</td>
<td>4.2</td>
<td>39.5</td>
</tr>
<tr>
<td>EMR</td>
<td>46.9</td>
<td>47.3</td>
</tr>
<tr>
<td>Rocchio</td>
<td>43.1</td>
<td>49.9</td>
</tr>
<tr>
<td>POP</td>
<td>46.3</td>
<td>55.7</td>
</tr>
<tr>
<td>HVIL (ours)</td>
<td>55.1</td>
<td>64.7</td>
</tr>
</tbody>
</table>

Human Feedback Analysis

- Boosting Rank-1 scores and pushing up true matches
- Saving annotation efforts

Conclusions

- A novel approach to human-in-the-loop person re-id by Human Verification Incremental Learning (HVIL) is proposed.
- HVIL model avoids the need for collecting off-line pre-labelled training data. It is able to learn cumulatively from human feedback.
- A regularised metric ensemble learning (RMEL) model is further developed to explore HVIL for automated re-id tasks when human feedback is unavailable.

References


Real-Time Monocular Segmentation and Pose Tracking of Multiple Objects

Henning Tjaden\textsuperscript{1}, Ulrich Schwanecke\textsuperscript{1} and Elmar Schömer\textsuperscript{2}

Overview

We present a real-time system capable of segmenting multiple 3D objects and tracking their pose using a single RGB camera, based on prior shape knowledge. The proposed method uses twist-coordinates for pose parametrization and a pixel-wise second-order optimization approach which lead to major improvements in terms of tracking robustness, especially in cases of fast motion and scale changes, compared to previous region-based approaches. Our C++/OpenGL implementation runs at 50–100 Hz on a commodity laptop when tracking a single object without relying on GPGPU computations.

In [1] we compare our method to the current state of the art [2] in various experiments involving challenging motion sequences and different complex objects.

Object\textsuperscript{1} is represented by a dense surface models consisting of points \( X_i = (x_i, y_i, z_i) \) in \( \mathbb{R}^3 \). The pre-calibrated and fixed intrinsic matrix of our camera denoted by \( \mathbf{K} \) and the pose of an object \( w_i \) denoted by \( \mathbf{T}_{cm,i} \), are given by

\[
\mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{T}_{cm,i} = \begin{bmatrix} t_{cm,i} & v_{cm,i} \\ 0 & 1 \end{bmatrix} \in \mathbb{SE}(3). \]

For pose optimization we represent the rigid body motion that occurred between two consecutive frames using twist

\[
\delta \theta = \begin{bmatrix} \omega & v \\ 0 & 1 \end{bmatrix} \in \mathbb{se}(3), \quad \text{with} \quad \omega = \begin{bmatrix} 0 \omega_2 & 0 \omega_3 & \omega_1 \omega_3 - \omega_2 \omega_0 \end{bmatrix} \in \mathbb{se}(3). \]

parametrized by \( \delta \theta = \theta \begin{bmatrix} w_1 \omega_2 & w_2 \omega_3 & w_3 \omega_1 \end{bmatrix} \), \( \theta \in \mathbb{R}^3 \), with \( w = (w_1, w_2, w_3, \theta_0) \), \( |\theta_0| = 1 \), where \( \theta \) is a one-parametric coupling of the rotation and translation parameters describing the motion along a screw. Each twist can be mapped to its corresponding rigid body transform via

\[
\exp(\delta \theta) = \exp(\omega \theta) \exp(v \theta) = \begin{bmatrix} 1 & -\omega_2 \theta & \omega_3 \theta & 0 \\ \omega_2 \theta & 1 & -\omega_1 \theta & 0 \\ -\omega_3 \theta & \omega_1 \theta & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \]

where \( \exp(\theta) \) can be computed according to Rodrigues’ formula. All images are undistorted non-linear distortion such that the perspective projection of a surface point to an image point is given by

\[
X_c = \pi(\mathbf{K} \mathbf{T}_{cm} X_0), \quad \text{with} \quad X_0 = (X/Z, Y/Z, Z). \]

A color image color is denoted by \( I : \Omega \rightarrow \mathbb{R}^3 \) where \( \Omega \subset \mathbb{R}^2 \) is the image domain. The color of each pixel \( x_i = (x, y) \in \Omega \) is given by \( y_i = I(x_i) \). Assuming pixel-wise independence we optimize the energy

\[
\mathcal{E} = \sum_{x \in \Omega} \mathcal{H}(\delta \Phi(x), \delta y) + (1 - \mathcal{H}(\delta \Phi(x), \delta y)) \delta y \]

directly for the pose parameters. Here \( \delta \Phi(x) \) and \( \delta y \) represent the foreground and background region membership probability per object of each pixel’s color (see Figure 1) and \( \delta \) is the level-set embedding of each object’s contour defined by its pose (see Figure 2). Thereby, the gradient of the energy is given by

\[
\frac{\partial \mathcal{E}(\delta \Phi, \delta y)}{\partial \Theta(\xi)} = \frac{1}{2} \left( \sum_{x \in \Omega} \frac{\partial \mathcal{H} (\delta \Phi(x), \delta y)}{\partial \Theta(\xi)} P_i(x) \delta y - P_i(x) \delta y \right) \delta \Theta(\xi), \]

where \( \Phi \) is the smoothed Dirac delta function corresponding to \( I_i \). Assuming small motion we get

\[
\frac{\partial \mathcal{E}(\delta \Phi, \delta y)}{\partial \Theta(\xi)} = \frac{1}{2} \left( \sum_{x \in \Omega} \frac{\partial \mathcal{H}(\delta \Phi(x), \delta y)}{\partial \Theta(\xi)} P_i(x) \delta y - P_i(x) \delta y \right) \delta \Theta(\xi), \]

where

\[
\mathcal{H}(\delta \Phi(x), \delta y) = \frac{1}{2} \sum_{x \in \Omega} \frac{\partial \mathcal{H}(\delta \Phi(x), \delta y)}{\partial \Theta(\xi)} P_i(x) \delta y - P_i(x) \delta y \]

and \( \mathcal{H}(\delta \Phi(x), \delta y) \)

\[
\sum_{x \in \Omega} \frac{\partial \mathcal{H}(\delta \Phi(x), \delta y)}{\partial \Theta(\xi)} P_i(x) \delta y - P_i(x) \delta y \]

is the corresponding rigid body transform and applied to the initial transform estimate as

\[
\mathbf{T}_{cm} \leftarrow \exp(\delta \Theta(\xi)) \mathbf{T}_{cm}. \]

To improve robustness and runtime, we compute this pose optimization in a coarse to fine manner.

Figure 3: Multiple object tracking with occlusion. Left: Common silhouette mask \( I_i \) of corresponding surface models \( m_1 \) and \( m_2 \). Right: \( \Phi^0 \) within \( h \leq 6 \) pixels around the occluded contour \( C^0 \) (grey values) and pixels influenced by occlusion (bright red inside, dark red outside of \( C^0 \)).

References


\textsuperscript{1}\textsuperscript{Computer Science Department, RheinMain University of Applied Sciences \textsuperscript{2}\textsuperscript{Institute of Computer Science, Johannes Gutenberg University Mainz}
**Problem statement**
Given an input motion sequence of oriented point clouds (with unknown correspondence) showing a dressed person, the goal is to estimate the body shape and motion of this person.

**Contributions**
- An automatic approach to estimate 3D human body shape in motion in the presence of loose clothing.
- A new benchmark consisting of 6 subjects captured in 3 motions and 3 clothing styles each that allows to quantitatively compare human body shape estimates.

**Method**
Overview of our pipeline:

Energy function:
\[
E = \omega_{data} E_{data} (\vec{F}, \vec{\beta}, \vec{\Theta}) + \omega_{cloth} E_{cloth} (\vec{F}, \vec{\beta}, \vec{\Theta})
\]

Optimization steps:
- Step 1. Pose initialization with Stitched Puppet [1] on every single frame.
- Step 2. Initial identity and pose estimation for the sequence.
- Step 3. Identity refinement based on the sequence considering wide clothing.

Clothing term:
\[
E_{cloth} = \sum_{i=1}^{N} \delta_{out} \delta_{NN} \left\| \vec{s}_i (\vec{\beta}, \vec{\Theta}) - N\hat{N} \left( \vec{s}_i (\vec{\beta}, \vec{\Theta}), \vec{F} \right) \right\|^2 + \omega_r \left\| \vec{\beta} - \vec{\beta}_0 \right\|^2
\]

Influence of \(E_{cloth}\) on walking sequence:

Left: input data overlayed with result with \(\omega_{cloth} = 0\) (left) and \(\omega_{cloth} = 1\) (right). Middle: cumulative per-vertex error of estimated body shape with \(\omega_{cloth} = 0\) and \(\omega_{cloth} = 1\). Right: color-coded per-vertex error with \(\omega_{cloth} = 0\) (left) and \(\omega_{cloth} = 1\) (right).

**Dataset**
A new dataset is acquired aiming to qualitatively evaluate body shape estimation algorithm. The dataset contains:
- 3 females and 3 males with large body shape variation.
- 3 motions: walk, knees-up, rotating upper body.
- 3 clothing styles: tight, layered, wide.
- Synchronized motion capture data and surface mesh.

Six representative examples:

**Result**
Example frames of layered clothing sequences

Example frames of wide clothing sequences

**Statistics:**

Comparison:

**References**

**Website**
Paper | Kinovis
A Shape-based Approach For Salient Object Detection using Deep Learning
Jongil Kim and Vladimir Pavlovic (jpkim and vladimir)@cs.rutgers.edu
Department of Computer Science, Rutgers, The State University of New Jersey

Motivation and Objective

- Visual saliency is one of the fundamental problems in computer vision
- Salient object detection identifies relevant parts of a visual scene
- Computational complexity may be reduced by filtering out irrelevant segments of the scene

- Salient object detection method using convolutional neural networks (CNNs)
- It aims to automatically identify the most important and salient regions/objects in the image
- CNN is used to estimate coarse saliency map of the target image
- We refined the estimated map using image specific low-to-mid level information

Examples for salient object detection

- Image
- Ground truth
- Our detection

Previous Approaches and Challenges

- Binary classification based approaches
  - The task in the binary classification is to estimate a saliency score of a pixel/patch using visual features extracted on around the pixel/patch
- Training the binary classifiers
  - A binary label is assigned to the patch based on the normalized overlap rate between the patch and its ground-truth salient map
  - The overlap rate ranges from 0 to 1; 0 represents no salient region and 1 means that the patch is fully contained in the salient region
  - The binary label is obtained by thresholding the overlap rate
- Limitations in the binary classification based approaches
  - They ignore the shape of the salient region
  - The patches whose overlap rates are around 0.5 are often ignored to prevent the classifiers from being confused
  - The valuable data is excluded from the training process
  - The geometric precision of saliency for overlaps around 0.5 will be significantly degraded

References


Proposed Method

Overview of the proposed method

- Input Image
- Region proposals
- Convolutional Neural Networks (CNN)
- Salient map by CNN
- Hierarchical segmentation
- Refined saliency map

Saliency representation

- Multi-label classification
  - \( X_k \): \( k \)-th region proposals
  - \( Y_k \): \( k \)-th encoded binary saliency map (\( m \times n \)) of region \( X_k \)
  - \( F_k(\ddot{X}_k) \): affine map from \( X_k \) to \( Y_k \)

Convolutional Neural Networks for salient object prediction

- The CNN consists of two branches to address both fine and coarse representations
- We tie weights between two branches up to fc7 so that the same fc-level feature embedding function is used for both coarse and fine patches

Refinement of saliency maps using hierarchical segmentation

\[
\text{Saliency map} = \sum_{i=1}^{L} w_i S_i
\]

Experimental Results

- Experimental settings
  - 4 widely used benchmark datasets, MSRA-5000, SOD, ECSSD, and PASCAL-S
  - 10 state-of-the-art methods in saliency detection, SF, HDCT, PCA, HS, MR, DFRI, LEGS, MB, DMC, and MDF
  - Evaluation methods: PR-Curve and F-Measure

- Experimental results

  - Recall
    - \( 0 \) to \( 1 \)
    - \( 0.1 \) to \( 0.9 \)
    - \( 0.2 \) to \( 0.8 \)
    - \( 0.3 \) to \( 0.7 \)
    - \( 0.4 \) to \( 0.6 \)
    - \( 0.5 \) to \( 0.5 \)
    - \( 0.6 \) to \( 0.6 \)
    - \( 0.7 \) to \( 0.7 \)
    - \( 0.8 \) to \( 0.8 \)
    - \( 0.9 \) to \( 0.9 \)
    - \( 1 \) to \( 1 \)

- Precision
  - \( 0 \) to \( 0.1 \)
  - \( 0.1 \) to \( 0.2 \)
  - \( 0.2 \) to \( 0.3 \)
  - \( 0.3 \) to \( 0.4 \)
  - \( 0.4 \) to \( 0.5 \)
  - \( 0.5 \) to \( 0.6 \)
  - \( 0.6 \) to \( 0.7 \)
  - \( 0.7 \) to \( 0.8 \)
  - \( 0.8 \) to \( 0.9 \)
  - \( 0.9 \) to \( 1 \)

- (Fine + Coarse representations) vs (Fine representation only)

  - GT
  - SSD-HS
  - SSD-HS

Conclusion

- We proposed a novel method to detect salient objects in an image using a convolutional neural network (CNN) model.
- Our method directly estimates the shape of the salient object using the CNN trained to predict the shape of the object
- We further refine the saliency map predicted by the CNN using the hierarchical segmentation maps to exploit the global information such as spatial consistency and object boundaries
Fast Optical Flow using Dense Inverse Search

Till Kroeger, Radu Timofte, Dengxin Dai, Luc Van Gool
Computer Vision Lab, ETH Zurich

Motivation

- **Optical Flow**: Find 2D displacement for each pixel in two-frame sequence.
- **Applications**: Low-level computer vision building block: Recognition, Tracking, Motion Segmentation, Robotic Navigation
- **Current focus in SOTA**: finding large displacements, learning flow, multi-view flow

Disregarding efficiency = Limited usefulness in practice!

Proposed method

- **Main contribution**: Optimal Flow at 600-300 Hz on single Core i7
- **Method overview**: For each scale (in coarse-to-fine pyramid) do:

  0) Create regular grid of overlapping 2D patches with fixed size and overlap.

  1) Estimate 2D displacement per patch. (Gradient descent on image intensities)

  \[ \Delta u = \arg \min_{u} \sum_{p} [I_{t+1}(x + u) - I_{t}(x)]^2 \]

  2) Weighted averaging of 2D patch displacements

  \[ U_{s}(x) = \frac{1}{Z} \sum_{b} \gamma_{b} U_{s}(x) \]

  3) Variational refinement

  \[ E(U) = \int \sigma(\Psi_{d}) + \gamma \Psi_{s}(E_{d}) + \alpha \Psi_{s}(E_{s}) \] 

- Excluding image preprocessing (image rescaling, gradients), IO (46-43 Hz with prep.)

Experimental results on MPI Sintel and KITTI

**MPI Sintel [1]**: Computer generated movie
Fast sideways motions
Occclusions, discontinuities

**KITTI [2]**: Realistic city, street scenarios
Forward motions, static scenes
Overexposure, Image noise

See paper for discussion of 1) operating points and parameter selection, 2) online benchmark results

Method comparison on MPI Sintel / KITTI

- **Ours**: Log. run-time (ms)
  - Time (ms)
  - Average. EPE

- **Others**: Log. run-time (ms)

High frame-rate optical flow

Solution for large displacement problem:
Instead of: \[ t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \rightarrow t_5 \]
Compute:
\[ t_1 \rightarrow t_2 \rightarrow t_3 \rightarrow t_4 \rightarrow t_5 \]

- Smaller displacements
- Reduced budget
- Accumulated flow errors / drift
- Increased image noise

Summary

- **Optical Flow at 600-300 Hz on single Core i7**
- **Code**: http://www.vision.ee.ethz.ch/~kroegert/OFlow/

Extensions, Future Work

- Implemented & Tested:
  - DIS for depth from stereo
  - Robust error norms (L1, Huber)
  - GPU Parallelization (OpenMP)

- Future work:
  - Sampling to recover small, fast object motions
  - GPU Parallelization
  - Exploration of more robust error metrics
  - Explicit reasoning of occlusions and motion borders

Acknowledgments: This work was supported by the European Research Council (ERC): VarCity (#273940) and the SNF: Tracking in the Wild (CRSII2_147693/1)

kroeger@vision.ee.ethz.ch http://www.vision.ee.ethz.ch/~kroegert/ http://www.vision.ee.ethz.ch/
1. Introduction

Multiview Registration

The goal of multiview registration is to find the rigid transformations \( M_{ij} \in SE(3) \) that bring multiple (\( n \geq 2 \)) 3D point sets into alignment.

We focus on global methods which consider simultaneously all the point sets: no drift

Our Contribution

We show that the registration problem can be expressed as a low-rank and sparse (LRS) matrix decomposition: robustness to outliers

2. Problem Definition

Multiview registration can be formulated as a motion averaging/synchronization problem without involving 3D points

\[
M_{ij} = M_{i} M_{j}^{-1} \quad \forall i, j \in \mathbb{L}
\]

where \( M_{ij} \in SE(3) \) is the rigid transformation that aligns point-set \( j \) with point-set \( i \).

Estimates \( \hat{M}_{ij} \in SE(3) \) are computed through the Iterative Closest Point (ICP) algorithm.

3. Proposed Approach

The synchronization constraint can be expressed as \( X = M M^{-1} \Rightarrow \text{rank}(X) = 4 \)

\[
X = \begin{pmatrix}
M_{11} & M_{12} & \cdots & M_{1n} \\
M_{21} & M_{22} & \cdots & M_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
M_{n1} & M_{n2} & \cdots & M_{nn}
\end{pmatrix},
M = \begin{pmatrix}
M_{1} \\
M_{2} \\
\vdots \\
M_{n}
\end{pmatrix},
M^{-1} = \begin{pmatrix}
M_{1}^{-1} & M_{2}^{-1} & \cdots & M_{n}^{-1}
\end{pmatrix}
\]

In the case of missing data, the available information is represented by \( P_{ij}(\hat{X}) \), where \( \hat{X} \) denotes a noisy version of \( X \), and \( P_{ij} \) denotes the projection onto the sampling set \( \Omega \).

Minimization Problem

\[
\min_{X, S} \| P_{ij}(\hat{X}) - X \|^2_F \\
\text{s.t. } X = M M^{-1}, M \in SE(3)^n, \sum_{i,j \in \mathbb{L}} \| \hat{M}_{ij} - M_{ij} \|^2_F,
\]

\[
\min_{X, S} \| P_{ij}(\hat{X}) - X \|^2_F \\
\text{s.t. } X = M M^{-1}, M \in SE(3)^n, S \text{ sparse in } \Omega
\]

This is a LRS decomposition: we aim at recovering the low-rank matrix \( L \) starting from an incomplete subset of its entries \( P_{ij}(\hat{X}) \) corrupted by noise \( N \) and outliers \( S \).

\[
P_{ij}(\hat{X}) = P_{ij}(L) + S + N
\]

The solution is then projected onto \( SE(3) \) via Singular Value Decomposition (SVD).

4. Low-rank and Sparse Decomposition

R-GoDec

R-GoDec \cite{arigoni2014robust} expresses the sparse term as \( S = S_1 + S_2 \) where \( S_1 \) has support on \( \Omega \) and represents outliers, while \( S_2 \) has support on \( \Omega^c \) and represents completion of missing entries.

\[
\min_{X, \lambda} \frac{1}{2} \| P_{ij}(\hat{X}) - L - S_1 \|^2_F + \lambda \| S_1 \|_1 + \lambda \| S_2 \|_p
\]

\[
\text{s.t. rank}(L) \leq 4, \supp(S_1) \subset \Omega, \supp(S_2) \subset \Omega^c
\]

References


Recognition from Hand Cameras: A Revisit with Deep Learning
Cheng-Sheng Chan*, Shou-Zhong Chen*, Pei-Xuan Xie*, Chiung-Chih Chang*, Min Sun*
National Tsing Hua University*

Wearable Camera System
Consisting of three cameras, two mounted on the left and right wrists to capture hands (referred to as HandCam) and one mounted on the head or chest (referred to as EgoCam).

Advantages about HandCam:
1. It avoids the need to detect hands and manipulation regions.
2. It observes the activities of hands almost all the time.

Goal
Using our system, we want to recognize user’s hand states including free vs. active (i.e., hands holding objects or not), object categories, and hand gestures.

We evaluate our system under two settings:
1. AllScene. Training and testing data share the same scenes.
2. Home. Test on a home scene, but trained in other scenes.

In total, we have collected data in three scenes: a large home, a medium-size lab, and a small office.

Hand vs. Egocentric
Egocentric view (top row) vs. Hand view (bottom row)

Here are some example images in our dataset:
[a, b] are the activities requiring hand-eye coordination. [a] is a standard example. [b] shows the case where HandCam is occluded. [c, d] shows the typical examples of missed hands in EgoCam but observed hands in HandCam.

Daily Life Dataset
We define a number of tasks, including 24 objects categories and 13 gestures, for users to follow. We asked 11 users to do these tasks by randomly selecting a task-order.

The dataset collected a new synchronized HandCam and EgoCam dataset with 20 videos captured in three scenes.

Model
The basic model is a 6-layer AlexNet, which used to recognize the hand-view or ego-view independently.

Next, we would like to combine HandCam with EgoCam to do even better. Intuitively, EgoCam should be complementary to HandCam in some ways.

Frame-based unary score $u(s_i)$.
The classifier generates a confidence $u(s_i)$ for state $s_i$ in $i^{th}$ frame. For example, $u(s_i=\text{Notebook})$ specifies the confidence that the $i^{th}$ frame contains a hand manipulating a notebook.

Change Detection
Since frame-based state recognition inevitably will contain spurious error predictions, we propose to detect possible state changes by training a frame-base binary state change classifier.

We propose the feature $c_i$, which measures semantic changes.

$$c_i = |f_{i+1} - f_i| = |f_{i+1} - f_{i-1}|,$$

where $f$ is the same deep feature used for state classifier.

Then we define a binary score function $b$ to calculate the score of switching state. Finally, we combine frame-based unary score with detected change candidates,

$$R(S) = \sum_{i=1}^{N} u(s_i) + \lambda \sum_{i=1}^{N-1} b(s_i, s_{i+1}) ,$$

$R(S)$ is the score as a function of a set of states $S = [s_1, s_2, ..., s_i, ...]$.

Indoor 3D reconstruction
To understand our dataset, we use the EgoCam to localize 6DoF camera poses. The localization results is overlaid on the 3D model of the scene constructed by a RGB-D camera. We visualize the camera trajectory from green (early in time) to red (late in time) dots. The good localization results suggest that EgoCam contain useful place information to help recognizing hand states.

Experiment Result
Quantitative.
HandCam’s performance is better than EgoCam in all setting. And Two-Streams model is better either HandCam or EgoCam.

Qualitative.

Unsupervised Visual Representation Learning by Graph-Based Consistent Constraints
Dong Li¹, Wei-Chih Hung², Jia-Bin Huang³, Shengjin Wang¹, Narendra Ahuja³, Ming-Hsuan Yang²
¹Tsinghua University, ²University of California, Merced, ³University of Illinois, Urbana-Champaign


Introduction

Supervised learning: Expensive annotations & poor scalability

Goal: Visual representation learning with a large, unlabeled image collection

Prior work:
Context [Doersch et al. ICCV’15]
Tracking [Wang and Gupta ICCV’15]

instance-level training data within the same image/video

Unsupervised Visual Representation Learning

Mining: Generate category-level training samples across different images

Cycle consistency → Positive mining
Geodesic distance → Negative mining

Training: Learn visual representations for binary classification

Experiments

Positive mining (CIFAR10)

Accurate positive pairs & Better CNN representations

Image classification (VOC’07)

Methods Supervision mAP
Agrawal ICCV’15 Ego-motion 52.9
Doersch ICCV’15 Context 55.3
Wang ICCV’15 Tracking triplet 58.4
SIFT+FV Matching pair 46.0
Ours Matching pair 56.5
Krizhevsky NIPS’12 Class labels 69.5

Competitive performance with the state-of-the-arts
Significant improvement over hand-crafted features

Semi-supervised learning (CIFAR10)

Boosted performance over directly fine-tuning

Image search (ImageNet)

Comparable to supervised learned representations
Introduction and Main Contributions

Main contributions:
- A new loss function for weakly-supervised semantic image segmentation.

Three loss terms: Seed, Expand and Constrain

\[ D = \{X, T\}_{i=1}^{n} \] - training data, where any X is an image and T is a set of its labels.
\[ \mathcal{C} \] - a set of all semantic categories.
\[ f_{\omega,c}(X; \theta) \] - probability score of class c at location u as predicted by deep convolutional neural network (CNN), f, parametrized by \( \theta \).

**Sec**: significantly outperforms previously proposed techniques.\[ \text{SEC (proposed)} \]

Experimental results

**Ablation study**

Effect of various loss terms.

```
Full loss
- seed + expand + constrain
  - seed
  - expand
  - constrain

**Training Data**

IST Austria Institute of Science and Technology

European Conference on Computer Vision, 19-22 September 2016, Amsterdam, Netherlands
1 Motivation

- The high-dimension-low-sample-size problem limits the flexibility of statistical shape models (SSMs).
- Collecting an adequately large and representative training population is often laborious and challenging, particularly in medical applications.
- Our approach: assumption of locality, i.e., we assume that local variations in shape, intensity or motion have limited effects in distant areas.
- Our method allows the model to combine local variations observed in different training samples while preserving overall object properties, i.e., generating valid instances.

2 Methods

Patch-based Sample Generation

- Random patches of different training shapes are combined in a virtual sample.
- Many virtual samples form the sparse data matrix $X \in \mathbb{R}^{m \times n}$.
- Patch sizes and patch distances depend on the desired level of locality.

Low-Rank Matrix Completion

- Solve the matrix-completion problem:
  $$M = \arg \min_M \|P_{\Omega}(X) - P_{\Omega}(M)\|_F^2 \quad \text{s.t. rank}(M) = k, \tag{1}$$
  
  with projection operator $P_{\Omega}(X)_{ij} = \left\{ \begin{array}{ll} X_{ij} & (i,j) \in \Omega \\ 0 & \text{else} \end{array} \right.$
- We use polar incremental matrix completion (PIMC) [1] to solve Eq. (1):
  – online algorithm based on GROUSE [2]
  – suitable for highly ill-conditioned data matrices $X$

Patch-based Model

- Solution of Eq. (1) is the matrix $M = UR^T$ of given rank $k$.
- Model basis $U$ is approximated to a new sample $y$ by
  $$y = Uw \quad \text{with} \quad w = \arg \min_w \|Uw - y\|_2^2, \tag{2}$$
- Distributions of $w \sim \mathcal{N}(\mu, \operatorname{Diag}(\sigma_1, \ldots, \sigma_n))$ are estimated from $w$.

3 Experiments

- Evaluation methods: Computing generalization errors and specificity errors for varying numbers of available training samples.
- Performance is demonstrated in three different applications:
  – IMM face data: 2D contour of 56 facial landmarks ($m = 116$) (see Fig. 3).
  – LIDC lung data: 3D surfaces with 2000 pseudo-landmarks ($m = 6000$).
  – Respiratory lung motion: 2D motion fields ($m = 64000$) (see Fig. 4).

4 Results

- A ROC-like analysis shows that our model clearly outperforms classical PCA and FEM-PCA models [3].
- In all applications the generalization error is improved for small training sizes (by $-20\%$ - $40\%$).
- As expected, improvements in terms of generalization ability come along with slightly increased specificity errors.

5 Conclusion

- The proposed method can be applied for a variety of problems.
- Leads to an increased flexibility and generalization ability while the validity of generated model instances is preserved.
- Online ability of the PIMC algorithm avoids explicit storage of data matrix $X$ in large scale problems, e.g., deformation modeling.
- Although only local information was provided, the proposed method is able to learn global shape variability as well (see Fig. 7).

References:
Motivation

- Visual recognition is a structured problem due to object-object, object-scene dependencies.
- Given an input $x$ and a set of tasks $Y = \{Y_t\}$, a chain model decomposes tasks sequentially:

$$P(Y=y|x) = P(Y_0=y_0|x) \cdot P(Y_1=y_1|x, y_0) \cdot P(Y_2=y_2|x, y_0, y_1) \ldots.$$

Chain Models in NLP

Seq2seq\(^1\) models use the chain rule to translate a source sentence to a target sentence.

Chain Models in Vision

- Single Image
  - The sequence of tasks is defined by the marginal distribution.
  - Weights are not tied in order to allow semantic flow of information.

- Video
  - The tasks are decomposed in time.
  - Weights are tied to enforce recurrence in time.

Scheduled Sampling\(^2\) is used to improve learning and bridge the gap between training and testing.

Chain Models for Pose Estimation

Single Image

- The task is to localize each joint in an image.
- The joints are decoded in descending order of accuracy: from easy to hard.
- Dataset: MPII Human Pose dataset.

<table>
<thead>
<tr>
<th>PCKh (%)</th>
<th>Torso</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>91.1</td>
<td>90.2</td>
<td>81.0</td>
<td>77.4</td>
<td>77.2</td>
<td>73.7</td>
<td>64.6</td>
<td>81.3</td>
</tr>
<tr>
<td>Chain Model</td>
<td>91.7</td>
<td>95.7</td>
<td>85.3</td>
<td>82.2</td>
<td>80.0</td>
<td>72.4</td>
<td>85.3</td>
<td></td>
</tr>
</tbody>
</table>

Error Analysis

Video

- The task is to track all joints across video frames.
- The joints are decoded in space and time: from easy to hard, from older to newer.
- Dataset: Penn Action dataset.

<table>
<thead>
<tr>
<th>PCK (%)</th>
<th>Head</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Model</td>
<td>94.1</td>
<td>90.3</td>
<td>84.2</td>
<td>83.5</td>
<td>88.7</td>
<td>87.2</td>
<td>87.7</td>
<td>87.5</td>
</tr>
<tr>
<td>Conv. RNN</td>
<td>95.3</td>
<td>92.5</td>
<td>87.9</td>
<td>87.5</td>
<td>91.1</td>
<td>89.8</td>
<td>90.1</td>
<td>90.1</td>
</tr>
<tr>
<td>Chain Model</td>
<td>95.6</td>
<td>93.8</td>
<td>90.4</td>
<td>90.7</td>
<td>91.8</td>
<td>90.8</td>
<td>91.5</td>
<td>91.8</td>
</tr>
</tbody>
</table>

---

\(^1\) Sutskever, Vinyals & Le. Sequence to sequence learning with neural networks, NIPS 2014

\(^2\) Bengio, Vinyals, Jaitly & Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks, NIPS 2015
Introduction

Motivation:
- Previous work shows improvement with better proposal methods [1]
- State-of-the-art CNN based action classification relies on multi-frame optical flow [2]
- Object recognition is improved by multi-region feature [3]

Contribution:
- We introduce a motion Region Proposal Network (RPN)
- We show that multi-frame optical flow significantly improves action detection
- We embed a multi-region scheme in the faster R-CNN model

Two-stream R-CNN for action detection

Pipeline of our Two-stream faster R-CNN (TS R-CNN)

Input: a single RGB frame and multi-frame optical flow

Training:
- RGB / appearance (A) and optical flow / motion (M) are trained separately with a VGG16 model pretrained on ImageNet.
- When training with multi-frame optical flow, we duplicate the VGG16 filters of the first layer multiple times and use the bounding boxes of the middle frame as ground truth.

Testing:
- Combine A & M models by developing a ReI fusion layer
- Average the softmax scores from A & M streams
- The bounding box regressors are applied to corresponding RoIs of each stream (the red and blue solid bars)

Multi-region two-stream faster R-CNN

Architecture of our multi-region TS R-CNN (MR-TS R-CNN)

- Introduce a multi-region generation layer next to the RPNs
- Introduce a mask-supported ReI pooling layer for border regions

Additional regions:
- Upper half, bottom half, border

Training: For both A & M streams, we fine-tune the network (only fully connected layer) of the original regions separately for each region

Linking and temporal localization

Given two regions $R_t$ and $R_{t+1}$ from consecutive frames, the linking score for class $c$ is

$$s_c(R_t, R_{t+1}) = s(a(R_t)) + s_b(R_{t+1}) + \beta \cdot \psi(ov)$$

where $\psi(ov)$ is the IOU overlap, $\beta$ is a scalar, $\psi(ov)$ is a threshold function

After linking, Viterbi algorithm is used to find optimal paths

Given a video-length detection $\mathcal{R}$, we solve the following objective to assign the start frame $s$ and end frame $e$,

$$s_c(\mathcal{R}(s,e)) = \arg\max_{s,e} \sum_{c \in \mathcal{C}} s_c(R_t - \lambda |L_c - L_c|)$$

$L_c$ is a class-specific mean length in training set

Action proposal evaluation

Comparison of different frame-level proposals on UCF-Sports (left) and J-HMDB (right)

- Non-tuned RPN obtains good proposals for actions
- Multi-frame optical flow RPN is better than single flow and saturates above 5 frames
- Best results are obtained by combining all the proposals of flow-5 and appearance RPNs

Experimental results

Frame-AP evaluation of TS R-CNN with varying number of frames and combination strategies on UCF-Sports (left) and J-HMDB split 1 (right)

- Multi-frame flow improves action detection and saturates above 5 frames
- Increasing the number of RGB frames does not improve the performance
- Our score fusion with a ReI fusion layer is better than a naive NMS method

Some actions are better represented by individual part R-CNN models
- Multi-region TS R-CNN model improves the Ours R-CNN model

Action detection examples with different region R-CNN models.

Reference

Semantic Co-segmentation in Videos

Yi-Hsuan Tsai1  Guangyu Zhong2  Ming-Hsuan Yang1
1UC Merced  2Dalian University of Technology

Code Available at https://sites.google.com/site/yihsuantsai/research/eccv16-cosegmentation

Introduction

• Goal
  – Segment objects in the video without prior knowledge
  – Understand visual semantics
• Motivation
  – Objects with the same semantics share more similarities
  – Solution: co-segmentation from a collection of videos

Algorithmic Overview

Video Collection

Semantic Tracklet Generation

Input Video

Semantic Object Clustering

Mean-shift

Object Proposals

Semantic Tracklets Generation

Remove noisy segments

Tracking segments to maintain temporal consistency

Semantic Tracklet Co-selection

• Goal
  – Discover true objects
  – Exploit relations between tracklets
• Questions
  – What is a good tracklet?
  – How to model relations?

Submodular Optimization for Tracklet Co-selection

• Submodular function
  – Model semantic similarity
  – Find high-quality object-like segments
• Pairwise term
  – Segments quality
  – Using similar segments
• Unary term
  – Use similar segments
  – Using similar segments

Quantitative Results

• Youtube-Objects dataset
  – 10 object categories
  – Baseline: frame-based method
• Other co-segmentation datasets
  – Input all the videos containing various objects

Qualitative Results

• Can segment multiple semantic objects under challenges such as fast movements, deformed shapes, occlusions and scale changes

Conclusions

• Propose an algorithm to co-segment objects and understand visual semantics from a collection of videos
• Videos with the same semantics share similarities and help each other
• Solve the co-segmentation problem by formulating the submodular function

2016 European Conference on Computer Vision (ECCV 2016)
1. Introduction

- Image recognition has been extensively studied in computer vision and machine learning.
- Inverse problem: generating images from high-level descriptions.

Contributions:
- Propose a novel problem of conditional image generation from visual attributes.
- Propose a novel layered foreground-background generative model based on conditional variational auto-encoders.
- Propose a general optimization-based method for posterior inference on novel images.

2. Attribute-conditional VAE

- Learn a model that generates image $x$ from given attribute $y$.

$$
\begin{array}{c}
\text{CVAE: Generation model} \\
\text{CVAE: Recognition model}
\end{array}
$$

Maximum likelihood of conditional distribution $p_{θ}(x|y)$ with stochastic variable $z$ is intractable.
Learning objective: variational lower bound of log-likelihood with an auxiliary recognition model $q_φ(\tilde{z}|x,y)$

$$
\log p_{θ}(x|y) = KL \left( q_φ(\tilde{z}|x,y)||p_θ(\tilde{z}|x,y) \right) + \mathcal{L}_{CVAE}(x, y; \theta, φ)
$$

where

$$
\mathcal{L}_{CVAE}(x, y; \theta, φ) = -KL \left( q_φ(\tilde{z}|x,y)||p_θ(\tilde{z}|x,y) \right) + \mathbb{E}_{q_φ(\tilde{z}|x,y)}[\log p_θ(x|y,z)]
$$

Layered Model: Disentangling CVAE

- Inspired by image composite model (Porter et al. SIGGRAPH 1984).
- Layered model: foreground layer, background layer and gating.
- Generate foreground image $x_F$, gating $g$ and background image $x_B$ with disentangled representation $z = [x_F, x_B]$

$$
x = x_F \odot (1 - g) + x_B \odot g
$$

3. Experiments

- LFW faces: 13,000 images from 6,000 identities with 73-dim score vector as attributes.
- CUB bird images: 12,000 images from 200 fine-grained bird species with 312-dim binary vector as attributes.
- Preprocessing: crop and resize the image to 64x64x3.
- Quantitative evaluation:
  - “Analysis by synthesis” on novel images
  - Posterior inference via optimization:
    $$p_θ(\tilde{z}|x,y) = \max_{z} \left[ \log p_θ(x|z,y) + \log p_θ(z|y) \right]$$
Referring expressions are sentences that identify a specific instance of an object.

Referring expressions rely on attributes and context for distinguishing one object from another.

Context between objects is usually specified using:
- spatial relationships such as “to the right”, “above”,
- interactions between objects such as “riding”, “holding”.

We address the problem of modeling context between regions to localize objects mentioned by referring expressions.

**Problem formulation**

The referred region is selected as the region with the highest likelihood of generating the referring expression.

\[
R^* = \arg \max_R p(S|R, I)
\]

**Baseline model**

The probability is modeled using an LSTM

**Word Embedding**

**Region CNN features**

**Region BBox**

**LSTM**

**Context region features**

**Context region BBox**

We then select the referred region by pooling the probabilities over all pairs using either the max function or the noisy-or function.

**Multiple Instance Learning (MIL)**

The challenge is that there is no annotation available for context objects mentioned in a referring expression.

Hence we construct positive and negative bags consisting of pairs of regions.

**Our model with pairwise context**

We consider \([\text{region, context\_region}]\) pairs when modeling the probability of the referring expression.

\[
p(S|R, R_c)
\]

**Training objectives with positive and negative bag margin**

We extend max-margin training objective of Mao et al. (CVPR 16) to margin based MIL objective functions.

\[
p(S|R, R_c) = \max_{\theta} \left\{ \log p(S|R, R_c) \right\}
\]

\[\frac{\lambda}{2} \min \left\{ \max_{\theta} \left\{ \sum_{i \in C} \left[ -\log p(S|R, R_i, R_c) \right] \right\}, M \right\}
\]

\[\frac{\lambda}{2} \min \left\{ \max_{\theta} \left\{ \sum_{i \in C} \left[ -\log p(S|R, R_i, R_c) \right] \right\}, M \right\}
\]

**Results**

<table>
<thead>
<tr>
<th>Google RefExp – Validation partition</th>
<th>Method \ Proposals</th>
<th>GT</th>
<th>MCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Likelihood [Mao et al]</td>
<td>57.5</td>
<td>42.4</td>
<td></td>
</tr>
<tr>
<td>Max margin [Mao et al]</td>
<td>65.7</td>
<td>47.8</td>
<td></td>
</tr>
<tr>
<td>Ours, Neg. Bag margin</td>
<td>68.4</td>
<td>49.5</td>
<td></td>
</tr>
<tr>
<td>Ours, Pos. &amp; Neg. Bag margin</td>
<td>68.4</td>
<td>50.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UNC RefExp – TestB partition</th>
<th>Method \ Proposals</th>
<th>GT</th>
<th>MCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Likelihood [Mao et al]</td>
<td>70.6</td>
<td>50.0</td>
<td></td>
</tr>
<tr>
<td>Max margin [Mao et al]</td>
<td>76.3</td>
<td>55.1</td>
<td></td>
</tr>
<tr>
<td>Ours, Neg. Bag margin</td>
<td>78.0</td>
<td>56.4</td>
<td></td>
</tr>
<tr>
<td>Ours, Pos. &amp; Neg. Bag margin</td>
<td>76.1</td>
<td>56.3</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>UNC RefExp – TestA partition</th>
<th>Method \ Proposals</th>
<th>GT</th>
<th>MCG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Likelihood [Mao et al]</td>
<td>65.9</td>
<td>53.2</td>
<td></td>
</tr>
<tr>
<td>Max margin [Mao et al]</td>
<td>74.9</td>
<td>58.4</td>
<td></td>
</tr>
<tr>
<td>Ours, Neg. Bag margin</td>
<td>75.6</td>
<td>58.6</td>
<td></td>
</tr>
<tr>
<td>Ours, Pos. &amp; Neg. Bag margin</td>
<td>75.0</td>
<td>58.7</td>
<td></td>
</tr>
</tbody>
</table>

Friction from Reflectance: Deep Reflectance Codes for Predicting Physical Surface Properties from One-Shot In-Field Reflectance

Hang Zhang¹, Kristin Dana¹, Ko Nishino²

¹Rutgers University, US  ²Drexel University, US

Introduction

One-shot in-field reflectance creates new opportunities for recognition and scene understanding. We present a model for reflectance patterns using a state-of-the-art representation that builds on key aspects of deep learning and binary embedding. We show that our approach of deep reflectance codes (DRC) provides an efficient representation of materials for recognition, for unsupervised material grouping and for the first of its kind friction-from-reflectance algorithm. Our novel framework uses angular reflectance patterns to go beyond recognition and estimate physical properties of the surface.

Reflectance and Friction Measurements

- Reflectance Disks
  - A parabolic mirror section is fixed so that its focus is at the surface point to be measured.
  - Pixel coordinates correspond to viewing angles.
  - Collimated white light passes through a movable aperture. The illumination direction is determined by the aperture position.

- Reflectance and Friction Database
  - 137 different materials were collected and grouped into 21 categories.
  - 2 spot and 3 exposure measurements per surface. 5 on-axis illumination angles with 2 different exposures.
  - The kinetic friction coefficients for each surface are measured

Methods

- Invariant Encoding
  - Angular gradient encodes discriminative information for material categories (Zhang et. Al. CVPR ’15).
  - Pre-trained CNNs can be used as feature extractors and Fisher Vector describes the pattern distributions and provides illumination invariant representation for reflectance disks.
  - The Fisher vector encoder (Perronnin et al. CVPR ’07) generalizes the BOW model to a probabilistic model.

- Deep Reflectance Codes
  - FV-CNN has achieved the state-of-the-art result for material recognition (Cimpoi et al. CVPR ’15).
  - FV-CNN is high-dimensional, the distance computation is expensive.
  - We learn a similarity preserving binary hash codes for fast image retrieval.

- Friction Prediction
  - Reflectance characterizes light interaction with a surface, encoding the intrinsic physical properties of the surface.
  - The DRC provides illumination invariance for our reflectance dataset and we show correspondence between the friction distribution and t-SNE embedded data.
  - We use the material reflectance representation to directly predict the friction coefficients by fitting a parameterized functional form.

Results

- Material Recognition
  - Comparison of encodings and hashing methods on different material datasets.
  - The randomized and optimized version of FV-CNN Hashing outperform the other methods in FTD, DTD and KTH datasets.

- Friction from Reflectance
  - (Left) t-SNE embedding of DRC of the reflectance disks. Classes are color-coded (21 classes). For some classes there is significant intra-class reflectance variation, but most group well within the t-SNE manifold.
  - (Right) Friction map generated in the t-SNE space.

- Scatter Plot of Friction Estimation
  - Friction prediction scatter plot of percentage error, colored by density.
  - The largest density of points have low prediction error. The overall mean percentage error is 12.05%.

Reference

http://ece.rutgers.edu/vision


This work was supported by National Science Foundation award IIS-1421134 to KD and KN and Office of Naval Research grant N00014-16-1-2158 (N00014-14-1-0316) to KN.
1. Introduction

1.1 Problem

- **Saliency priors**, which are shown to be effective in previous work, are completely discarded by most CNN based methods.
- A **limited size** of local image patches is considered by most CNNs methods.
- **Binary classification** problems, which represent saliency detection, have relatively weak supervision information.

1.2 Our Solution

- Incorporation for the saliency priors into the network to facilitate training and inference.
- The **recurrent structure** to refine the coarse inference from previous time steps.
- A **RFCN pre-training method** for saliency detection using semantic segmentation data to both leverage strong supervision from multiple object categories and capture the intrinsic representation of generic objects.

2. Overview

- Employs three kind of low-level contrast features, including color, intensity and orientation, and the center prior knowledge to introduce **saliency prior maps**.
- Train the RFCN with two stage training strategy, pre-training on the segmentation data set and fine-tuning on the saliency data set. The recurrent structure can incorporate the saliency prior maps into the CNNs with an end-to-end training method.
- Refine the saliency maps with a post-processing method which can improve the performance of RFCN. Edge-preserving maps are produced with the computation of color confidence and spatial confidence.

3. Algorithm

3.1 Saliency Prior Maps

- We encode prior knowledge into a saliency prior map which serves as the input to the network. The priors include color, intensity and orientation feature contrast. Then we integrate these priors together and filter the result with a gaussian function, which proves center prior information.

\[
P(s) = u(s) \times (g(s) + I(s) + O(s))
\]

3.2 Recurrent Architecture

- Architecture 1: Forward propagation of the whole network is conducted in every time step, which is very expensive in terms of both computation and memory. We adopt this architecture for accuracy in our paper.
- Architecture 2: In the t-th time step, the predicted foreground map in the last time step serves as saliency prior map. The deconvolution part takes the convolution feature map as well as the foreground map to refine the saliency prediction.

3.3 Training RFCN

- Pre-training is conducted on the PASCAL VOC 2010 semantic segmentation data set. Saliency detection and semantic segmentation are highly correlated but essentially different in that saliency detection aims at separating generic salient objects from background, whereas semantic segmentation focuses on distinguishing objects of different categories.
- After pre-training, we modify the RFCN network architecture by removing the first C+1 channels of the last feature map and only maintaining the last two channels.

3.4 Post-Processing

- Given the final saliency score map predicted by the RFCN, we first segment the image into foreground and background regions by thresholding it with its mean saliency score.
- We then compute a spatial confidence and a color confidence score for each pixel
- Finally, we weight the predicted saliency scores by spatial and color confidences to dilate the foreground region.

4. Results

5. Conclusions

- We propose a recurrent fully convolutional network based saliency detection.
- Our method integrates low level saliency prior knowledge and fully convolutional neural networks with a recurrent structure.
- Experimental results on five benchmark data sets show that the proposed algorithm achieves favorable results against the state-of-the-art methods.
We compare Deep3D against baselines by measuring pixel-wise reconstruction loss:

We also conduct human subject experiments to evaluate the quality of stereo pairs generated by different algorithms. In each trial we show the subjects two 3D image and instruct them to pick the one with better 3D effect. The table below shows the chance of the row methods being picked over the column methods:

Incorporate lower level features with learned upsampling.
In contrast to traditional methods, Deep3D does not need expensive ground-truth depth maps for training. Instead, it’s directly trained on existing 3D movie frames.
We achieve this by estimating a probabilistic disparity map $D$. A selection layer then combines $D$ with the original input $I$ to generate the final output $O$:

$$O_{ij} = \sum_d d_{ij}^D D_{ij}$$

where $d$ ranges over all disparity levels, $\sum_d D_{ij}^D = 1$ for all $i,j$ and $D_{ij} = i_{ij,d}$.
During testing, we upsample the soft disparity map instead of directly upsample the final prediction for higher image quality.

Deep3D automatically converts 2D images or videos to stereo 3D, for viewing with 3D glasses or VR displays.

In our experiments, we find that Deep3D outperforms baselines with and without oracle mean disparity.

We also conduct human subject experiments to evaluate the quality of stereo pairs generated by different algorithms. In each trial we show the subjects two 3D image and instruct them to pick the one with better 3D effect. The table below shows the chance of the row methods being picked over the column methods:

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Disparity</td>
<td>7.75</td>
</tr>
<tr>
<td>Eigen et. al.</td>
<td>7.75</td>
</tr>
<tr>
<td>Deep3D</td>
<td>6.87</td>
</tr>
<tr>
<td>Eigen et. al. + Oracle</td>
<td>6.31</td>
</tr>
<tr>
<td>Deep3D + Oracle</td>
<td>5.47</td>
</tr>
</tbody>
</table>

Oracle means the ground truth image-wise mean disparity is given. We observe that Deep3D outperforms baselines with and without oracle mean disparity.

Deep3D is developed with MXNet, a flexible, scalable, and memory efficient deep learning package. Our code for the experiments can be downloaded at https://github.com/piiswrong/deep3d
Problem and Contributions

Problem

- Existing approaches:
  - Neglected the problem of adapting the features or the learned model over time
  - Manual labeling of supervised approaches require huge human effort

Contributions

- Low-rank similarity-dissimilarity metric learning method which
  - introduces sparsity inducing regularizers that allow identification and exploitation of the most discriminative dimensions for matching and
  - can be trained in an incremental fashion by means of a stochastic Alternating Directions Methods of Multipliers (ADMM) optimization algorithm
- A method to reduce the human labeling effort required to properly update the model
- An unsupervised graph-based approach identifies the most relevant gallery persons for every single probe

Approach Overview

Low-Rank Sparse Similarity-Dissimilarity Learning

- Image feature representations $x \in \mathbb{R}^d$ might be very high-dimensional and contain non-discriminative components
- We propose to learn a low-rank similarity function and a low-rank dissimilarity measure which self-determines the discriminative dimensions of the underlying manifold

Objective

i. Learn a similarity function and a dissimilarity metric
\[
\sigma_K(x_P, x_G) = x_P^T K x_G
\]
P \in \mathbb{R}^{d \times d}
\[
\delta_p(x_P, x_G) = \|P x_P - PX_G\|_2^2 = \|x_P - x_G\|^T P^T P (x_P - x_G)
\]
ii. Let the score function be
\[
S_{K,F}(p,g) = (\Omega_K + \Omega_F) / (\Omega_K + \Omega_F)
\]

Model Adaptation with Reduced Human Effort

- Perform incremental steps to minimize the augmented Lagrangian with new image pairs that are progressively acquired as time passes
- This requires human labeling of such pairs
- To limit the manual effort and improve model generalization, we aim to select only a small set of informative gallery persons to update the model

Probe Relevant Set

i. Represent the probe-gallery set for a single person as an undirected graph $G$ with no loops
ii. Cluster $G$ in such a way that
   a. A cluster contains the probe and gallery persons which are similar to each other
   b. All persons outside a cluster should be dissimilar to the ones inside

Experimental Results

- 3 Benchmark Datasets
- ViPeR (90’ viewpoint variations)
- PRID450s (background clutter and occlusions)
- Market1501 (1,501 persons and 6 cameras)
- LOMO feature descriptor
- Batch-update scheme
  - 1st batch for model initialization
  - 2nd-4th batches to simulate updates

Human Effort

- Unsupervised vs Supervised Probe-Gallery Set Selection
- Fixed vs Dynamic Probe-Gallery Set Selection
- State-of-the-art Comparisons
Image Quality Assessment Using Similar Scene as Reference

Yudong Liang, Jinjun Wang, Xingyu Wan, Yihong Gong, and Nanning Zheng
Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University

Overview

Contributions:
- The first work to support IQA from reference image with similar scene but is not aligned.
- A Dual-path deep Convolutional Neural Network (DCNN)
- The state of the art performance

Exist work
- Full reference image quality assessment (FR IQA)
- No reference image quality assessment (NR IQA)
- Reduce reference image quality assessment (RR IQA)
- Learning based methods
- Human: Comparing vs making direct judgement

Motivations
- The Non-existence problem of reference images:
  - Wide availability of images
  - Not aligned
- Studies in Human Visual System (HVS):
  - Spatial frequency, luminance, structural information
  - The visual attention property (saliency), a smaller but representative part of the scene
- Learning based methods
  - Learn discriminative feature for IQA

Proposed Method

Experimental Results

2.1 Training sample generation
- No public datasets for our problems

Experimental Results

2.2 experiments
- Three datasets: Live and Tid2008 (select 4 distortions) Collected by us: in-house dataset
- Linear Correlation Coefficients (LCC) and Spearman Rank Order Correlation Coefficient (SROCC)

The Influence of structural similarity
- The influence of geometrical transformation
- IQA referenced with different structural similarities

Conclusions and Future work
- Image quality assessment using non-aligned reference images with similar scene is solvable.
- A Dual-path deep CNN model handle this problem well.
- Exploration for IQA similar reference images selection, collecting larger non-aligned IQA dataset and building deep model of different architecture.
Overview

- Facial attributes have implicit latent correlations – which are naturally learnt through a multi-label approach.
- Due to demographic correlations, it is difficult to obtain a balanced multi-label facial attribute training set; thus a multi-label classifier trained for one demographic might perform poorly on another due to training set biases.
- CelebA, for example, contains images of celebrities where Young is over-represented and anti-correlated attributes like Bald and Gray Hair are under-represented.

Our Approach

- Idea: Backpropagate multi-label Euclidean loss, weighting frequency discrepancy between source distribution $S_i$ and target distribution $T_i$.
- Given binary labels $\{+1, -1\}$, let $S_i^+$ and $S_i^-$ be the number of occurrences in the source distribution and $T_i^+$ and $T_i^-$ be the number of occurrences in the target distribution.
- The domain adaptive weights for each class then become:

$$P(i|+1) = \begin{cases} 1 & \text{if } T_i^+ > S_i^+ \\ \frac{S_i^+}{T_i^+} & \text{otherwise} \end{cases} \quad \text{and} \quad P(i|-1) = \begin{cases} 1 & \text{if } T_i^- > S_i^- \\ \frac{S_i^-}{T_i^-} & \text{otherwise} \end{cases}$$

The mixed-objective Euclidean loss function across an $N$-sample data tensor $X$ with $N \times M$ label matrix $Y$ and classification output $f(\cdot)$ is then:

$$L(X,Y) = \sum_{j=1}^{N} \sum_{i=1}^{M} P(i|Y_{ji})\|f_i(X_j) - Y_{ji}\|^2.$$ 

Using the CelebA training partition, we trained a VGG-16 base architecture, replacing the final layer with our mixed-objective loss function.

Re-Balanced CelebA: CelebAB

- Re-balancing the classifier to a uniform target distribution yields a shift in output scores distributions.

Re-balancing yields superior attribute-based verification rates on the LFW dataset.

Evaluation: CelebA Dataset Un-Balanced

- For evaluation on the same dataset, we assume that $S_i = T_i$.
- MOON advances the state of the art over the Features+Classifiers approach with a relative reduction in classification error of 28.7%.
- Our approach also outperforms separate end-to-end networks of the same topology.

Approaches to the Balancing Problem

- Features + Classifiers (FaceTracer, L Nets + ANet): Derive a feature space from non-attribute face data with attribute classifiers trained in that feature space.
  - Easy to balance. Requires less training data.
    - Assumes that attribute related features will be implicitly embedded in the feature space with no explicit guidance.
  - Separate End-To-End DCNNs (Separate): Train one network per attribute label; hope that the network learns correlations between attributes without explicitly optimizing to that end.
    - More accurate.
    - Leads to a huge representation. Requires more data.
- Mixed-Objective Optimization Network (MOON): Combine the objectives of multi-label optimization with re-balancing under a domain-adaptive mixed objective function.
  - Most accurate. Allows for a compact, domain-adaptive, multi-label representation.
  - Requires a new approach to balancing (this work).

Mixed-Objective Optimization Network (MOON): A Mixed Objective Optimization Network for the Recognition of Facial Attributes

Features + Classifiers

MOON DCNN(based on VGG16)

Separate End-To-End DCNNs

 проблемы с чтением изображений
Degeneracies in Rolling Shutter SfM

Cenek Albl1 Akihiro Sugimoto2 Tomas Pajdla1

CMP, FEE, Czech Technical University in Prague1 National Institute of Informatics, Japan2

KEY CONCEPTS

• SFM with rolling shutter (RS) images is not straightforward
• RS Camera motion models introduce degeneracies

• Degeneracy – images explained by a distorted scene not representing reality
• Readout direction – a direction in the scene in which the image lines read out
• GS BA – Bundle adjustment with standard perspective camera model
• RS BA – Bundle adjustment that incorporates some of the RS camera models
• General unordered sets of images – images captured by cameras where constraints on the relative camera motion between images cannot be imposed

WHY ARE THERE DEGENERACIES?

Rolling shutter projection
\[ \alpha_i \omega_i = \alpha_i [c_i, r_i, t_i]^T = R_x(r_x)R_y c_0 + c_0 + r_x t \]

\[ R_x(r_x) = \begin{bmatrix} 1 & r_x \omega_x & r_x \omega_y \\ -r_x \omega_x & 1 & -r_x \omega_z \\ -r_x \omega_y & r_x \omega_z & 1 \end{bmatrix} \]

\[ \omega_x = \omega_x = -0.3 \quad \omega_y = -0.6 \quad \omega_z = -1 \]

Using rotation around the axis we can collapse the camera and explain the scene by arbitrarily flat scene in the direction.

When \( \omega_x = -1 \) and \( \omega_y = -1 \) all back-projected 3D points lie in a single plane.

Hedborg et. al. 2012

Albl et. al. 2016

HOW TO AVOID DEGENERACIES

• The further the directions are from being parallel, less flattening will occur.

• When readout directions between 3 or more cameras are more than 30 degrees apart, the scene is reconstructed correctly.

DEGENERACIES IN RS SFM OBSERVED

• RS bundle adjustment flattens the scene in one direction, why?

• A single degeneracy prevents correct reconstruction in most cases
• Guide how to capture data to avoid this degeneracy is provided
Deep Deformation Network for Object Landmark Localization
Xiang Yu, Feng Zhou and Manmohan Chandraker
NEC Laboratories America

Introduction
Our objective:
- Consistent localization for semantic landmarks
- Unified architecture for various tasks
- Demonstration on face, human body and bird

Figure 1: Limitation of cascaded-CNN for various tasks.

Limitations of prior works:
- Cascaded architecture is task-specific
- Computationally expensive on cascading

Figure 2: Our unified structure for multiple tasks.

Our approach:
- SBN provides regularized initialization
- PTN captures local deformation
- Same network structure for multiple tasks
- No cascades, very fast runtime, 4ms on GPU

Figure 3: Shape Basis Network.

SBN predicts the shape as the initialization of PTN using the PCA assumption.

\[ y_s = y + Q x_s, \quad \text{where } x_s = f(w_s, x). \] (1)

Localization Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ESR</th>
<th>SDM</th>
<th>ERT</th>
<th>LBF</th>
<th>cGPRT</th>
<th>DDN (Ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-W</td>
<td>7.58</td>
<td>7.52</td>
<td>6.40</td>
<td>6.32</td>
<td>5.71</td>
<td>5.65</td>
</tr>
</tbody>
</table>

Table 1: Mean relative error (%) on 300W.

Human Body Part Localization:

<table>
<thead>
<tr>
<th>Method</th>
<th>Head</th>
<th>Shoulder</th>
<th>Elbow</th>
<th>Wrist</th>
<th>Hip</th>
<th>Knee</th>
<th>Ankle</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pishchulin et al.</td>
<td>87.2</td>
<td>56.7</td>
<td>46.7</td>
<td>38.0</td>
<td>61.0</td>
<td>57.5</td>
<td>52.7</td>
<td>57.1</td>
</tr>
<tr>
<td>Tompson et al.</td>
<td>90.6</td>
<td>79.2</td>
<td>67.9</td>
<td>63.4</td>
<td>69.5</td>
<td>71.0</td>
<td>64.2</td>
<td>72.0</td>
</tr>
<tr>
<td>Chen &amp; Yuille</td>
<td>91.8</td>
<td>78.2</td>
<td>71.8</td>
<td>65.5</td>
<td>73.3</td>
<td>70.2</td>
<td>63.4</td>
<td>73.4</td>
</tr>
<tr>
<td>DDN (Ours)</td>
<td>87.2</td>
<td>88.2</td>
<td>82.4</td>
<td>76.3</td>
<td>91.4</td>
<td>85.8</td>
<td>78.7</td>
<td>84.3</td>
</tr>
</tbody>
</table>

Table 2: PCK at 0.2 on LSP dataset.

Bird Part Localization:

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>Methods</th>
<th>Ba</th>
<th>Be</th>
<th>By</th>
<th>Bt</th>
<th>Cn</th>
<th>Fo</th>
<th>Le</th>
<th>Li</th>
<th>Lw</th>
<th>Na</th>
<th>Re</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>Ning et al.</td>
<td>9.4</td>
<td>12.7</td>
<td>8.2</td>
<td>9.8</td>
<td>12.2</td>
<td>11.3</td>
<td>7.8</td>
<td>6.7</td>
<td>11.5</td>
<td>12.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>18.8</td>
<td>12.8</td>
<td>14.2</td>
<td>15.9</td>
<td>16.2</td>
<td>20.3</td>
<td>7.1</td>
<td>8.3</td>
<td>13.8</td>
<td>19.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.05</td>
<td>Ning et al.</td>
<td>46.8</td>
<td>62.5</td>
<td>40.7</td>
<td>45.1</td>
<td>59.8</td>
<td>63.7</td>
<td>66.3</td>
<td>33.7</td>
<td>31.7</td>
<td>54.3</td>
<td>63.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>66.4</td>
<td>49.2</td>
<td>56.4</td>
<td>60.4</td>
<td>61.0</td>
<td>60.0</td>
<td>66.9</td>
<td>32.3</td>
<td>35.8</td>
<td>53.1</td>
<td>66.3</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: PCK at 0.02 and 0.05 on CUB200-2011.
Learning Visual Storylines with Skipping Recurrent Neural Networks
Gunnar A. Sigurdsson, Xinlei Chen, Abhinav Gupta

- We learn the visual and temporal parts of concepts: Wedding, Christmas
- Learn in an unsupervised way with recurrent networks
- Our S-RNN learns the common long-term latent stories
- We train one model for each concept (e.g., Paris)

Method

- Train RNN over all possible subsequences
- Rewritten as sequential update equations for E and M steps
- Fits in the RNN pipeline: Infer next variable, then backprop
- Prediction is softmax over future images
  - Images are represented as their fc7 features

RNN/LSTM

compared to

S-RNN

Storylines

"Wedding"  

Using the probabilistic model Find the likeliest sequence

Finding a image sequence that captures a concept

Summarization

Use latent variables, infer subsequence

Finding a short summary of a photo album
Inferring latent variables gives a summary
Album can be summarized in various ways:

Prediction

Using the model sequentially Predicting what comes next

Predicting what comes next in a photo album
Towards large-scale city reconstruction from satellites
Liuyun DUAN and Florent LAFARGE
INRIA Sophia Antipolis, France

Problem Statement

- Producing compact 3D city models from stereo pairs of satellites
- Retrieving geometry and semantics simultaneously (robustness to occlusion and low image quality)
- Operating at the scale of geometric atomic regions (preservation of urban object shapes, scalability and performance)

Key Ideas

- Operating at the scale of geometric atomic regions
- Retriving geometry and semantics simultaneously
- Robustness to occlusion and low image quality
- Scalability and performance

Our method

1. Partition images into connected convex polygons by [4] and estimate their elevation by SGM [5]
2. Label simultaneously the semantic classes and the elevations of each polygonal partition

\[ U(f) = \sum_{l \in P} D_{data}(l) + \sum_{(c, l) \in C} V_{semantic}(l, c) + \sum_{(c, l) \in C} V_{coupling}(l, c) \]

- data term: measures the coherence between the elevation estimates of a polygon and its proposed label
- smooth term: penalizes \( E_s \) adjacent polygons with different labels proportionally to their radiometric similarity
- coupling term: penalizes \( E_c \) adjacent polygons with different labels proportionally to their overlapping ratio

3. Re-label elevations into the horizontal plane

- 3 types of cells when projecting polygons into the horizontal plane:
  - coherent cells: inherit two identical elevations from left and right partitions
  - conflict cells: inherit at least one elevation, and not coherent cells
  - empty cells: no inherited elevation

Re-labeling conflict and empty cells by energy minimization with elevations \( \{e_0, e_1, \ldots, e_n\} \)

- data term: (i) empty cells prefer to be ground \( e_0 \), and (ii) conflict cells to be roof with an elevation value as close as possible to its inherited elevations.
- pairwise potential: penalizes adjacent cells with different labels whose common edges do not back-project well into input images

Experiments

<table>
<thead>
<tr>
<th>Customer</th>
<th>Building reconstruction</th>
<th>Geometric accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denver, US</td>
<td>100 MB</td>
<td>100 M</td>
</tr>
<tr>
<td>Seoul, South Korea</td>
<td>0.35M</td>
<td>8.5 min</td>
</tr>
<tr>
<td>Melbourne, Australia</td>
<td>0.5M</td>
<td>13.7 min</td>
</tr>
<tr>
<td>Alexandria, Egypt</td>
<td>1.35M</td>
<td>32.2 min</td>
</tr>
</tbody>
</table>

References

Weakly Supervised Object Localization Using Size Estimates
Miaojing Shi and Vittorio Ferrari

Task
Weakly supervised object localization
input: motorbike

Method

Size Estimator
- Input: 4096D CNN image features
- Output: size \( \sqrt{S} \); \( \sigma \)
- Method: kernel ridge regression
- Class-specific regressor
- Train: PASCAL VOC 12 trainval
- Test: PASCAL VOC 07 trainval

Inter-batch order
Recall = \( \frac{|Q^c_{GT} \cap Q^c_{ES}|}{|Q^c_{GT}|} \)

Class - Chair

Results on PASCAL VOC 07 (20 classes)

MIL Baseline*
+ Size Order
+ Size Weighting

Ground Truth

* AlexNet features + Linear SVM + Objectness [Dollar ECCV14]

Class – Chair

Train on trainval, measure CorLoc; Test on test, measure mAP

Method

<table>
<thead>
<tr>
<th></th>
<th>CorLoc</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>39.1</td>
<td>20.1</td>
</tr>
<tr>
<td>Our</td>
<td>46.3</td>
<td>24.9</td>
</tr>
<tr>
<td>Scheme</td>
<td>55.8</td>
<td>28.0</td>
</tr>
<tr>
<td>Baseline</td>
<td>60.9</td>
<td>36.0</td>
</tr>
<tr>
<td>Cimbis PAMI16</td>
<td>43.2</td>
<td>24.7</td>
</tr>
<tr>
<td>Wang TIP15</td>
<td>48.5</td>
<td>31.6</td>
</tr>
<tr>
<td>Blen CVPR15</td>
<td>43.7</td>
<td>27.7</td>
</tr>
</tbody>
</table>

- Both size order and size weight improve results
- Full system outperforms state-of-the-art

Size Order
- Current way: training on the entire set at the same time
- Our way: curriculum learning using object size estimates
  - Size Order
    - Big (easy)
    - Small (hard)
  - Size Weighting

Object detection

Size Weighting
- Size weighting function
- Estimation order

Weakly supervised object localization
input: motorbike

Current way: training on the entire set at the same time

Both size order and size weight improve results

Full system outperforms state-of-the-art
Supervised Transformer Network for Efficient Face Detection

Dong Chen, Gang Hua, Fang Wen, Jian Sun
Microsoft Research

Introduction

• Contributions
  • A new state-of-the-art face detector
  • End-to-end trained cascaded network
  • Supervised Transformer Layer automatically learns the optimal canonical location
  • Non-top K Suppression achieves better recall than NMS
  • ROI convolution scheme speeds up 3x on with little recall drop
  • Achieve best result on FDDB, AFW and PASCAL faces datasets

• Challenge Samples

Overview

Steps:
1. Multi-task Region Proposal Network (RPN)
   • Produces a set of candidate face regions along with landmarks
2. Non-top K Suppression
   • Keep top K candidate face regions in a local neighborhood
3. Supervised Transformer Layer
   • Warp the face regions into a canonical pose
4. Multi-granularity R-CNN
   • Joint features from RPN and R-CNN network

Experiments

• Comparison with the State-of-the-art

• Results on FDDB, AFW and PASCAL faces datasets

Algorithm Details

• Supervised Transformer Layer
  • Motivation: Alignment is helpful to distinguish faces/non-faces
  • Problem: Canonical position selection
  • Our solution:
    Model as function: $\hat{I} = f(I, x, y, \tilde{x}, \tilde{y}, \tilde{l})$
    Update with gradient: $\frac{\partial L}{\partial x}, \frac{\partial L}{\partial y}$

• Forward formula:
  $\begin{bmatrix} a_i - m_x \\ b_i - m_y \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} m_x \\ m_y \end{bmatrix}$

• Solution:
  $a = \frac{c_1}{c_3}, b = \frac{c_2}{c_3}$

• Backward formula
  • Supervision information from classification objective of R-CNN network
  $L = \sum_{(i,j)} \frac{\partial L}{\partial (x,y)} \frac{\partial (x,y)}{\partial (\tilde{x}, \tilde{y})}$

• Non-top K Suppression
  • Make detector practical – CNN only calculates regions may contain faces

ROI convolution

• ROI convolution pipeline

Challenge Samples
A Geometric Approach to Image Labeling
Freddie Åström*,†, Stefania Petra*,‡, Bernhard Schmitzer† and Christoph Schnörr*§

*HCI, †MIG, ‡IPA, Heidelberg University, Germany
§CEREMADE, University Paris-Dauphine, France
freddie.astroem@iwr.uni-heidelberg.de

Introduction

We introduce a smooth non-convex approach to image labeling in a novel geometric framework which complements established approaches to image labeling.

- The major underlying concept is a smooth manifold of probabilistic assignments of a pre-specified set of prior data (the "labels") to given image data.
- The Riemannian gradient flow, with respect to a corresponding objective function, evolves on the manifold and terminates, for any \( \delta > 0 \), within a \( \delta \)-neighborhood of an unique assignment (labeling).
- Our approach is numerically implemented with sparse, highly-parallel interior-point updates that efficiently converge, largely independent from the number of labels.

Method

- **The Assignment Manifold.** The relative interior \( S = \text{int} \{ W \} \) of the probability simplex \( \Delta_n = \{ p \in \mathbb{R}^n : \sum_i p_i = 1 \} \) becomes a differentiable Riemannian manifold when endowed with the Fisher-Rao metric:

\[
\langle (p, q) \rangle = \sum_{i} p_i q_i, \quad \forall p, q \in T_S \Delta_n, \quad T_p S = \{ v \in \mathbb{R}^n : \langle v, \mathbf{1} \rangle = 0 \}, \quad p \in S.
\]

**Definition 1 (Assignment Manifold).** The manifold of assignment matrices, called assignment manifold, is the set

\[
\mathcal{W} = \{ W \in \mathbb{R}^{n \times n} : W_{ij} > 0, \quad i \in [m], \quad j \in [n] \}.
\]

- **Distance Matrix, Lifting Map and Riemannian Means.** The Riemannian distance between a prior set \( P_T \) and the data \( J \) is

\[
D = \mathbb{E}_{p \in \Delta_n} \mathbb{E}_{q \in \Delta_n} \mathbb{E}_{f \in P_T} \mathbb{E}_{j \in [n]} D_{ij} \langle p, q \rangle, \quad \rho > 0, \quad i \in [m], \quad j \in [n].
\]

Given \( D \) and \( W \), we lift the vector field \( D \) to the manifold \( \mathcal{W} \) by

\[
L = L(W) = \exp_{W}([-U]) \in \mathcal{W}, \quad U_{ij} = D_{ij} - \frac{1}{\rho} \langle D, U \rangle, \quad i \in [m].
\]

Based on the likelihood matrix \( L \), we define the similarity matrix

\[
S = S(W) \in \mathcal{W}, \quad S_{ij} = \exp_{W}([-L])_i \mathbb{E}_{j \in [n]}
\]

**Definition 2 (Lifting Map (Manifolds \( S, \mathcal{W} \))).** The lifting mapping is defined by

\[
\exp : T_W \mathcal{W} \rightarrow \mathcal{W}, \quad (W, U) \mapsto \exp_{W}(U) = \left( \begin{array}{c} \exp_{W}(U_{1,1}) \\ \vdots \\ \exp_{W}(U_{m,n}) \end{array} \right)
\]

where \( U_{i,j} \in [m] \) index the rows of the matrices \( U, \mathcal{W} \).

- **Objective Function.** The assignment matrix \( W \) in \( \mathcal{W} \) can be thought of as posterior probabilities,

\[
W_{ij} = P(f_i | f_j),
\]

of assigning prior feature \( f_i^* \) to the observed feature \( f_j \). A natural objective function to be maximized is

\[
\max_{W \in \mathcal{W}} J(W), \quad J(W) = \langle S(W), W \rangle.
\]

The functional \( J \) together with the feasible set \( \mathcal{W} \) formalizes the following objectives:

1. Assignments \( W \) maximally correlate with the feature-induced similarities \( S = S(W) \)
2. Assignments of prior data to observations should be done in a spatially coherent way.
3. Maximizers \( W^* \) should define image labels in terms of rows \( W^*_{j,k} = e^k \in \{0,1\}^n, \quad i, k \in [m], \) that are indicator vectors.

**Lemma 1.** Let \( \overline{\mathcal{W}} \) denote the closure of \( \mathcal{W} \). We have

\[
\sup_{W \in \mathcal{W}} J(W) = \sup_{W \in \overline{\mathcal{W}}} J(W) = m,
\]

and the supremum is attained at the extreme points

\[
\overline{\mathcal{W}} = \{ W \in \{0,1\}^{m \times n}, \quad W_{i,k} = e^k, \quad i \in [m], \quad k_1, \ldots, k_m \in [n] \} \subset \overline{\mathcal{W}},
\]

corresponding to matrices with unit vectors as rows vectors.

Experiments

- **Parameter Influence on Labeling.** The labeling problem comprises of \( 31 \) color vectors forming the prior data set \( P_T \) with the distance \( \|f_i - f_j\| = \|f_i - f_j\|_1 \):

\[
|\mathcal{W}| = 3 \times 3, \quad |\mathcal{W}| = 5 \times 5, \quad |\mathcal{W}| = 7 \times 7.
\]

- **Convergence.** Left. Small spatial scales necessitate to resolve more conflicting assignments through propagating information by geometric spatial averaging. Right. Higher selectivity leads to faster convergence.

- **Image inpainting by Labeling.** The selectivity parameter was set to \( \rho = 1 \) and the neighborhood size was \( 5 \times 3 \).

Further results


A Minimal Solution for Non-Perspective Pose Estimation from Line Correspondences

Gim Hee Lee (gimhee.lee@nus.edu.sg)
National University of Singapore

Objective

Given:
1. Three 2D-3D line correspondences \( l_{i}, \leftrightarrow L_{W}, \forall i = 1, 2, 3. \)
2. \( l_{i} \) are seen respectively by three cameras \( F_{C_{i}}, \forall i = 1, 2, 3 \) and \( L_{W} \) is defined in a fixed world frame \( F_{W}. \)
3. \( F_{C} \) are rigidly fixed together with known camera intrinsics \( K_{i} \) and extrinsics \( T_{C_{i}W}, \forall i = 1, 2, 3. \)

Find: The pose of the multi-camera system with respect to the fixed world frame, i.e., relative transformation \( (P_{W}^{C_{1}}, t_{W}^{C_{1}}) \) that brings a point defined in the multi-camera reference frame \( F_{C} \) to the fixed world frame \( F_{W}. \)

Fig.1. An illustration of the non-perspective pose estimation problem from line correspondences.

Note: In general, the minimal case for non-perspective pose estimation can be three 2D-3D line correspondences from a multi-camera system made up of either two or three cameras.

NP3L: 3-Line Minimal Solution

Solving for \( R_{W}^{C_{1}} \)

- Since \( U^{T}V \) is zero for any Plücker line \( [U^{T} V^{T}]^{T}, u_{C} \) is parallel to \( U_{C} \) and \( R_{W}^{C_{1}} = R_{W}^{C_{2}} R_{W}^{C_{3}}, \) from \( L_{C} \leftrightarrow l_{C} \) we get
  \[ u_{C}^{T} R_{W}^{C_{2}} R_{W}^{C_{3}} R_{W}^{C_{1}} U_{W} = 0. \]  
  (2)
- Rewrite Eq. (2) into \( ar = 0 \), \( a \) is a 1 x 9 matrix from known variables \( u_{C}, R_{W}^{C_{2}} \) and \( V_{W} \), and \( r \) is a 9-vector from the unknown rotation matrix \( R_{W}^{C_{1}}. \)
- Given three 2D-3D line correspondences in the minimal case, we set \( ar = 0 \) where \( A \) is a 3 x 9 matrix from \( a_{i}, \forall i = 1, 2, 3. \) Right null-space of \( A \) gives
  \[ r = \beta_{1} b_{1} + \beta_{2} b_{2} + \beta_{3} b_{3} + \beta_{4} b_{4} + \beta_{5} b_{5} + \beta_{6} b_{6}. \]  
  (3)
- Setting \( \beta_{6} = 1 \) and enforcing the orthogonality constraint on the rotation matrix formed by \( r \), we get a system of 10 polynomial equations in terms of the unknowns \( \beta_{1}, ..., \beta_{5}, \)
  \[ f_{j}(\beta_{1}, \beta_{2}, \beta_{3}, \beta_{4}, \beta_{5}) = 0, \quad j = 1, 2..., 10, \]  
  (4)
which can be solved with Gröbner basis to get eight solutions.

Solving for \( t_{W}^{C_{1}} \)

- Since \( U_{C} \) and \( u_{C} \) are parallel, from \( L_{C} \leftrightarrow l_{C} \) we get
  \[ \lambda u_{C} = (R_{W}^{C_{2}} R_{W}^{C_{3}} | R_{W}^{C_{2}} t_{W}^{C_{2}} + t_{W}^{C_{2}} | R_{W}^{C_{3}} R_{W}^{C_{1}}) \begin{pmatrix} U_{W} \\ V_{W} \end{pmatrix}. \]  
  (5)
- Taking cross-product of \( u_{C} \) on both sides to get rid of the unknown scalar value \( \lambda \) and with three 2D-3D line correspondences, we get \( Bt = 0. \)
- Solution of the unknown \( t = [t_{W}^{C_{2}} t_{W}^{C_{3}} t_{W}^{C_{1}}]^{T} \) is given by the right null-space of \( B \) which is made of known variables variables \( R_{W}^{C_{2}}, t_{W}^{C_{2}}, R_{W}^{C_{3}}, u_{C}, U_{W} \) and \( V_{W}. \)

NPnL: n Line Correspondences

- For \( n \geq 3 \) 2D-3D line correspondences, \( A \) and \( B \) are \( n \times 9 \) and \( 2n \times 4 \) matrices. The solution steps remain the same as the minimal case.

Special Cases

- One Camera: becomes the perspective pose estimation problem where all line correspondences are seen by only one camera. Camera extrinsics \( (R_{W}^{C_{2}}, t_{W}^{C_{2}}) \) vanishes and we directly solve for the camera pose \( (R_{W}^{C_{1}}, t_{W}^{C_{1}}) \) without any change to the algorithm.
- Parallel 3D Lines: minimal case where two or all the three lines are parallel. Rank of matrix \( A \) drops below 3 and \( R_{W}^{C_{3}} \) cannot be solved. Fortunately, we can prevent this degenerate case by omitting parallel lines.

Plücker Line Representation

- 2D-3D line correspondence \( l_{C} \leftrightarrow L_{W}. \)
- \( P_{W}^{C_{1}}, P_{W}^{C_{2}}, \) end-points of 3D line \( L_{W}. \)
- \( p_{a}, p_{b}: \) end-points of 2D line \( l_{C}. \)

Fig.2. A 2D-3D line correspondence represented as Plücker lines.

- Plücker representation of a 3D line segment in the world frame is a 6-vector \( L_{W} = [U_{W}^{C_{1}} V_{W}^{C_{1}}] \) computed from the 3D line end-points \( P_{W}^{C_{1}}, P_{W}^{C_{2}}. \)
- \( L_{W} \) can be expressed in the camera reference frame \( F_{C}. \)

\[ L_{C} = (U_{C}^{T} V_{C}^{T}) = (T_{W}^{C_{i}} L_{W} \in \begin{pmatrix} R_{W}^{C_{i}} & t_{W}^{C_{i}} \end{pmatrix}) L_{W}. \]  
(1)
- \( R_{W}^{C_{1}} = R_{W}^{C_{2}} R_{W}^{C_{3}} \) and \( t_{W}^{C_{1}} = R_{W}^{C_{2}} t_{W}^{C_{2}} + t_{W}^{C_{2}}, \) where \( (R_{W}^{C_{2}}, t_{W}^{C_{2}}) \) is the known camera extrinsics and \( (R_{W}^{C_{3}}, t_{W}^{C_{3}}) \) is the unknown pose of the multi-camera system.
- Plücker representation of the 2D line correspondence is a 6-vector \( l_{C} = [u_{C}^{T} v_{C}^{T}]^{T} \) computed from the 2D line end-points \( p_{a}, p_{b}. \)

Reprojection Error for RANSAC

\[ \begin{pmatrix} P_{C_{1}}, P_{C_{2}} \end{pmatrix}: \) end-points of 3D line transformed into camera frame.
- \( p_{a}, p_{b}: \) reprojection of \( P_{C_{1}}, P_{C_{2}} \) onto the image.
- \( p_{a}, p_{b}: \) camera matrix normalized 2D line end-points.

Reprojection Error:
\[ e = \frac{d_{a} + d_{b}}{2(|p_{a} - p_{b}|)}. \]

Fig.3. An illustration of reprojection error from a 2D-3D line correspondence.

Results

Simulations (Non-Perspective):

Real Data (Non-Perspective):

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Cameras</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>3</td>
</tr>
<tr>
<td>02</td>
<td>5</td>
</tr>
<tr>
<td>03</td>
<td>10</td>
</tr>
<tr>
<td>04</td>
<td>20</td>
</tr>
<tr>
<td>05</td>
<td>25</td>
</tr>
</tbody>
</table>

Dataset 02

Simulations (Perspective):

Real Data (Perspective):

Table 1: Average error from the multi-camera systems simulated with 3, 5, 10, 20 and 25 cameras from datasets 01 and 02 respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of Cameras</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>3</td>
</tr>
<tr>
<td>02</td>
<td>5</td>
</tr>
<tr>
<td>03</td>
<td>10</td>
</tr>
<tr>
<td>04</td>
<td>20</td>
</tr>
<tr>
<td>05</td>
<td>25</td>
</tr>
</tbody>
</table>

Dataset 02

Table 2: Comparisons of errors from our algorithm, Zhang et. al. and Mirzaei et. al. for one camera.

References:

Zhang et. al.: Robust and efficient pose estimation from line correspondences, ACCV 2012.
Mirzaei et. al.: Globally optimal pose estimation from line correspondences, ICCRA 2011.
Natural Image Stitching with the Global Similarity Prior

Yu-Sheng Chen and Yung-Yu Chuang

Abstract
This paper proposes a method for stitching multiple images together so that the stitched image looks as natural as possible. A comprehensive evaluation shows that the proposed method consistently outperforms several state-of-the-art methods, including AutoStitch, APAP, SPHP and ANNAP.

Method
Mesh Optimization
\[ \tilde{V} = \arg\min_V \psi_a(V) + \psi_l(V) + \psi_g(V) \]

1. Alignment Term
   - ghost removal
2. Local Similarity Term
   - shape preserving
3. Global Similarity Term
   - naturalness

Rotation selection
\[ \theta_i = (u_i, v_i) \]

Relative Rotation Range
Minimum Line Distortion Rotation

Scale selection
\[ s_i = \frac{f_0}{f_i} \]

2D Method
\[ E_{MLDR} + \lambda_E E_{ZERO}, \text{ where} \]
\[ E_{ZERO} = \sum_{i \in \Omega} \left\| \begin{bmatrix} u_i \\ v_i \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\|^2 \]
\[ E_{MLDR} = \sum_{(i,j) \in \mathcal{E}} \left\| R(\phi_{ij}) \begin{bmatrix} u_i \\ v_i \end{bmatrix} - \begin{bmatrix} u_j \\ v_j \end{bmatrix} \right\|^2 \]

3D Method
\[ E_{MLDR} + \lambda_E E_{ALPHA}, \text{ where} \]
\[ E_{ALPHA} = \sum_{(i,j) \in \mathcal{O}} \left\| R(\alpha_{ij}) \begin{bmatrix} u_i \\ v_i \end{bmatrix} - \begin{bmatrix} u_j \\ v_j \end{bmatrix} \right\|^2 \]
\[ E_{MLDR} = \sum_{(i,j) \in \mathcal{O}} \left\| R(\phi_{ij}) \begin{bmatrix} u_i \\ v_i \end{bmatrix} - \begin{bmatrix} u_j \\ v_j \end{bmatrix} \right\|^2 \]

An example of stitching six images

Comparison with other methods:
- AutoStitch
- SPHP + APAP
- AANAP
- Ours

AANAP
Ours
Ours(2D)
Ours(3D)

Compose the focal length by bundle adjustment with better initialization and point matches.
Minimal Solvers for Generalized Pose and Scale Estimation from Two Rays and One Point

Federico Camposeco¹, Torsten Sattler¹ and Marc Pollefeys¹,²
¹ETH Zürich, Switzerland
²Microsoft

Summary and Contributions
- A camera VO/SfM trajectory can be seen as a generalized camera.
- Getting the absolute pose of such camera is known as gPnP or gPnP+s.
- These trajectories must have local 3D points; therefore we propose:
  - New formulations of the gPnP and gPnP+s where one 3D point is known:
    - A gPnP+s solver in closed form (quartic) instead of octic (~50x faster)
    - A gPnP solver in closed form (quadratic) instead of octic (~50x faster)

Geometric Solutions
- **Unknown Scale: 1 Point 2 Rays (gPnP+s)**
  - Use ratio of distances
    \[ \frac{p_i - p_{i+1}}{p_i - p_{i+1}} \approx \frac{-D_{i-1}}{D_i} \]
  - Solve a single quartic polynomial (in closed form)
    \[ a_1 \lambda^4 + a_2 \lambda^3 + a_3 \lambda^2 + a_4 \lambda + a_5 = 0 \]
- **Known Scale: 1 Point 2 Rays (gPnP)**
  - Define sphere-ray distance
    \[ d_i(\lambda_i) = (f_i(\lambda_i) - D_i, t)^2 \]
  - Minimize it w.r.t depth
    \[ \frac{\partial d_i}{\partial \lambda_i} = 0, \text{ for } i = 1, 2, 3 \]
  - By solving two independent quadratics
    \[ \lambda_i (N_i + [q_i x q_i]) = D_i, t \]

Methods Compared
- gPnP Chen: Earliest solution to gPnP iterative with up to 16 solutions.
- gPnP Kneip: Solves the rotation of (B) directly. Returns 8 solutions.
- gPnP Lee: Same point-ray representation as ours. Returns 8 solutions.
- gPnP+s: Returns up to 8 solutions, from 4 point-raymatches.
- gDLs: 3D point alignment via SVD of at least 3 points.

References
6 Zhang, R. Camera Calibration with Multiple Cameras. TR. 1998.

Simulation Results
- Numerical Stability
- Effects of Measurement Noise and Quality of Triangulation
- Timing Comparison (C++ implementations)

Real Data Results
- Absolute Orientation
- SfM and Query Trajectory

This project was funded by Google's Project Tango.
Learning to Hash with Binary Deep Neural Network
Thanh-Toan Do, Anh-Dzung Doan, Ngai-Man Cheung
Singapore University of Technology and Design
{thanthoan_do, dung_doan, ngaiman_cheung}@sutd.edu.sg

Hashing for Image Search
- Nearest neighbor search in database of $N$ data points in $\mathbb{R}^D$
  - Exact search: $O(ND)$ in both time and space → difficult for handling large scale, e.g., million SIFT features in 3D model.
  - Finding nearest neighbors in Hamming space more efficient
  - Time and space: $O(NL)$ instead of $O(ND)$
  - Hamming distance by hardware operation, i.e., XOR.

Contributions
- Define hash function as a deep network
- In order to deal with binary constraints, instead of involving the sign or step function as in previous work, our design constrains one layer to directly output the binary codes
- Propose a framework to solve the resulting NP-hard optimization with binary constraints
- Directly incorporate the independence and balance properties on the codes
- Ensure the similarity preserving
- Binary Deep Neural Network for both unsupervised and supervised hashing

Unsupervised Hashing

\[
\min_{W,c,B} \frac{1}{2m} \|X - W^{(1)}B - c^{(1)}1_m\|_2^2 + \frac{\lambda_1}{2} \|H^{(1-1)}B - I\|_2^2 + \frac{1}{2} \sum_{l=1}^{n-1} \|W^{(l)}\|_2^2
\]

s.t. $B \in \{-1,1\}^{m \times n}$

- $1^{st}$ term: reconstruction error → encourage neighbor preserving
- $3^{rd}$ term: binary quantization loss
- $4^{th}$ and $5^{th}$ terms: independence and balance of codes

Optimization
- Alternative optimize over network weight $(W,c)$ and $B$.
  - When fixing $B$: unconstrained opt w.r.t $(W,c)$: solving by LBFGS
  - When fixing $(W,c)$: binary opt w.r.t $B$
    - Unsupervised: coordinate descent, i.e., solve one row of $B$ at a time → closed form for that row.
      Let $V = X - c^{(1)}1_mQ = (W^{(1)}B - c^{(1)}1_m)Q + \lambda_1H^{(1-1)}$. For $k = 1 \ldots L$. Let $w_k$ be the $k^{th}$ column of $W^{(l)}$. $w_k$ be matrix $W^{(l)}$ excluding $w_k$. $q_k$ be the $k^{th}$ column of $Q^T$. $b^l_k$ be the $k^{th}$ row of $B$. $b^l_k$ be matrix of $B$ excluding $b^l_k$
      We have closed form for $b^l_k = sgn(q_k^Tw_k) w_k^T W_k B_k$.
    - Supervised: closed form $B = sgn(H^{(1)}1_m)$

Dataset
- CIFAR10: 60K images; 10K testing; AlexNet features
- MNIST: 70K images; 10K testing; raw (intensity) features
- SHIFTM: 1M features; 10K testing; SIFT features

Evaluation of Unsupervised Hashing

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10</th>
<th>MNIST</th>
<th>SIIF1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>UH-BDNN</td>
<td>0.55-0.79</td>
<td>0.24-0.32</td>
<td>0.08-0.15</td>
</tr>
<tr>
<td>BA</td>
<td>0.55-0.65</td>
<td>0.28-0.32</td>
<td>0.08-0.17</td>
</tr>
<tr>
<td>ITQ</td>
<td>0.5-0.65</td>
<td>0.28-0.32</td>
<td>0.08-0.17</td>
</tr>
</tbody>
</table>

Evaluation of Supervised Hashing

<table>
<thead>
<tr>
<th>Method</th>
<th>CIFAR10</th>
<th>MNIST</th>
<th>SIIF1M</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>SH-BDNN</td>
<td>0.55-0.79</td>
<td>0.24-0.32</td>
<td>0.08-0.15</td>
</tr>
<tr>
<td>SDH</td>
<td>0.55-0.79</td>
<td>0.24-0.32</td>
<td>0.08-0.15</td>
</tr>
<tr>
<td>ITQ</td>
<td>0.5-0.65</td>
<td>0.28-0.32</td>
<td>0.08-0.17</td>
</tr>
</tbody>
</table>

Comparison with CNN-based hashing methods:
- CIFAR10: 50K training; 10K testing (with leave-one-out procedure)
- MNIST: 70K images; 10K testing
- SH-BDNN achieves more than 3% improvement over the state of the art.

Conclusion
- Unified framework based on deep neural network for both unsupervised and supervised hashing.
- Directly produce binary code at one layer.
- Ensure good properties of codes: similarity preserving, independence, balance.
- Code release: http://tinyurl.com/hx4f44f
Automatically Selecting Inference Algorithms for Discrete Energy Minimisation

Paul Henderson • Vittorio Ferrari

**Problem**
MAP inference in graphical models (GMs)
many algorithms...
GM characteristics affect which is best
we automatically select best algorithm for a given GM

**Prediction Tasks**
- best-and-fastest
- good-and-fastest

**Baselines**
- naive
  - lowest energy
- strong
  - use best for superclass (pairwise, higher-order, partitioning)

**Algorithms**
- LP dual
- combinatorial
- AD
- Gurobi-LP
- FastPD
- ICN
- TRW-S
- QPBO
- Parallel LBP
- Serial LBP
- Gurobi-ILP
- A*
- move-making
- \(\alpha\)-expansion
- ICM
- message-passing

**Dataset**
- 340 GMs from 32 problem classes
- OpenGM2 + four others
- 50:50 train / test split
- 340 GMs

**Features**
- instance size
- structural features
- energy features
  - fractions of pairwise factors...
  - (i) are submodular
  - (ii) satisfy metric axioms
  - influence of each factor order on final energy

**Accuracy**
- mean fraction of variables matching best labelling
- fraction of GMs correctly classified

**Confusion Matrix**
- \(\alpha\)-exp
- BPS
- FPD
- ICM
- KL
- LBP
- QPBO
- TRW-S
- UM

**Speed**
- speed-up
- matching labels

- run every algorithm and take the best
  - 1x
  - 100%

- predict then run selected best-and-fastest algorithm
  - 47x
  - 97.1%

- predict then run selected good-and-fastest algorithm
  - 88x
  - 96.4%

- labelings very similar to running every algorithm
- much lower computational expense
Ego2Top: Matching Egocentric Viewers in Egocentric and Top-view videos

Shervin Ardebor
University of Central Florida
Orlando, FL
ardesir@cs.uchicago.edu

Ali Borji
University of Central Florida
Orlando, FL
aborji@crce.ucf.edu

We explore the relationship between egocentric and top view videos, by matching and finding correspondences between them. Having a set of egocentric videos, and a top-view video, we aim to answer two main questions:

**Question 1:** Does the top-view video contain these egocentric viewers? i.e. are the camera holders present in the top-view video?

- Given a set of top-view videos, we model each with a graph and evaluate them.
- We rank the graphs based on the graph matching similarity score of their graph and the egocentric graph, and measuring the ranking performance.
  - **Blue curve:** performing graph matching only using the nodes (setting all edge similarities of the affinity matrix to zero).
  - **Red curve:** incorporating both nodes and edges in the graph matching score.

**Question 2:** Knowing a top view video contains the egocentric viewers, can we tell which viewer is capturing each of the egocentric videos?

- **Assignment Accuracy:** Percentage of egocentric videos that have been matched to the correct identity in the top-view video. (Evaluating hard-assignment)
  - Baselines:
    - Rnd: Random Assignment
    - #F: Only using the 1D unary features
    - G-F: Using the 2D unary features
    - G-M: Using both unary and binary features and performing graph matching.
- **Ranking Performance:** For each egocentric video, we rank the top-view viewers based on their soft-assignment probabilities and evaluate the rank of the correct match. (Evaluating soft-assignment)
  - In both evaluations, incorporating the edges has the most contribution due to edges outnumbering nodes and therefore providing more information.

**Graph Matching:**

- Affinity matrix $A$ is formed between the two graphs:
  - Diagonal elements: $A_{ii} = \max_i (U^T_i U_i + (1 - \alpha) \max_j (U_j^T U_j))$.
  - Non-diagonal elements: $A_{ij} = \max (B^T_{ij} B_{ij})$.
- Graph Matching is defined as: $\arg \max x^T A x$.

Where $x$ is a binary vector containing the assignments, and $A$ is an affinity matrix.

- A spectral graph matching method is used to compute a soft-assignment between the nodes of the two graphs by performing eigen-decomposition on $A$.
- Soft-to-hard assignment is done using the hungarian method to come up with a binary vector $x$, including the assignments.
- The graph matching score is used as a measure of the similarity of the egocentric set and the top view video. Sorting different top view videos based on that score answers the first question, and the hard assignment output answers the second question.

**Relative graph size vs Performance**

Assignment and ranking performance was evaluated using different subsets of the egocentric set. Left shows the ranking and assignment accuracy only using unary features, and right is using unary and pairwise in spectral graph matching.
Cross-modal Supervision for Learning Active Speaker Detection in Video

Punarjay Chakravarty and Tinne Tuytelaars

Abstract

In this paper, we show how to use audio to supervise the learning of active speaker detection in video. Voice Activity Detection (VAD) guides the learning of the vision-based classifier in a weakly supervised manner. The classifier uses spatio-temporal features to encode upper body motion - facial expressions and postures - associated with speaking. We further improve a generic model for active speaker detection by learning person specific models. Finally, we demonstrate the online adaptation of generic models learnt on one dataset, to previously unseen people in a new dataset, again using audio (VAD) for weak supervision. The use of temporal continuity overcomes the lack of clean training data. We are the first to present an active speaker detection system that learns on one audio-visual dataset and automatically adapts to speakers in a new dataset. This work can be seen as an example of how the availability of multi-modal data allows us to learn a model without the need for supervision, by transferring knowledge from one modality to another.

Applications

- Video conferencing
- Human – Robot interaction
- Video diarization

Spatio-temporal Feature Extraction

Improved Dense Trajectories (IDT) are extracted from within upper body tracks, pooled with Fisher vectors and used as features for the active speaker classifier.

Weak Supervision from Voice Activity Detection

Voice Activity Detection (VAD) based audio supervision tells us someone is speaking. But not who. Structured Output Learning is used to learn a latent SVM classifier in the presence of partially observed (latent) inputs.

Audio weakly supervises Video Learning

- Voice Activity Detection (VAD)
- Improved Dense Trajectories
- Fisher vector pooling
- Upper body detection
- Tracking
- Latent SVM (offline)
- Generic classifier
- Weighted SVM (online)
- Person-specific classifier

Improvements with Online Learning on the Columbia Dataset

The Columbia dataset is a panel discussion of 5 speakers over 35 minutes. 2-3 people are in frame at any one time, and only one of them is speaking. The generic classifier learnt offline on other data improves with online learning using just a few seconds of video on specific speakers in the Columbia video. Again, audio VAD weakly supervises online learning.

Normalized raw scores (blue) with the online learnt classifier and thresholded and temporally smoothed speak/non-speak values for speakers Sick (red) and Long (green), along with Ground Truth (GT, solid colour = speak), in minutes 27:00 to 40:00 in the Columbia dataset.
Problem: Semantic Pixel Labeling in a Video

Key Idea: Account for Local and Long-range Spatiotemporal Cues Using Deep Learning

Contributions

- New deep architecture — Recurrent Temporal Deep Field (RTDF).
- New efficient mean-field inference that jointly predicts hidden variables in RTRBM and CRF and labels pixels in videos.
- New end-to-end joint training of all components of RTDF.

Formulation

$$E_{RTDF}(y^t, h^t | y_{<t}, I') = \Sigma_p \psi_1(x^t_p, y^t_p) - \Sigma_{p,p'} \psi_2(y^t_p, y^t_{p'})$$

Mean-field Inference

$$\bar{y}^t = \arg\max_{y^t} \Sigma_h \exp(-E_{RTDF}(y^t, h^t | y_{<t}, I'))$$

Acknowledgment

This work was supported in part by grant NSF RI 1302700.
Conventional face super-resolution methods, also known as face hallucination, become very fragile when the input low-resolution image size is too small and upscaling factors are large. To address these shortcomings, we present a discriminative generative network that can ultra-resolve a very low resolution face image of size 16×16 pixels to its 8x larger version by reconstructing 64 pixels from a single pixel. We introduce a pixel-wise L2 regularization term into the network and backpropagating its residual, it is possible to ultra-resolve in any size while GANs can only generate images in fixed and small sizes.

When training our network, we only require frontal and approximately aligned images, which makes the magnification task even more challenging as almost all facial details are missing. To the best of our knowledge, our method is the first attempt to develop discriminative generative networks for generating authentic face images. We demonstrate that our UR-DGN achieves better visual results than the state-of-the-art. When using a pixel-wise L2 regularization term into the network and backpropagating its residual, it is possible to ultra-resolve in any size while GANs can only generate images in fixed and small sizes.

We show that by introducing a pixel-wise L2 regularization term into the network and backpropagating its residual, it is possible to ultra-resolve in any size while GANs can only generate images in fixed and small sizes.

When training our network, we only require frontal and approximately aligned images, which makes the training datasets more attainable. Our UR-DGN can ultra-resolve regardless of pose, lighting and facial expression variations.

Due to its feed-forward topology, our ultra-resolution method is very fast.

Objective function:
\[ \min_{D,G} F(G,D) = E_{x \sim p_{data}}[\log D(h_x)] + E_{z \sim p_{z}}[\log(1 - D(G(z)))] + \lambda E_{z \sim p_{z}}||h_z - h_x||_2^2 \]

Algorithm 1 Minibatch stochastic gradient descent training of UR-DGN

Inputs: minibatch size N, LR and HR face image pairs \((x, h)\), maximum number of iterations \(K\).
1. while \( \text{iter} < K \) do
2. Choose one minibatch of LR and HR image pairs \((x, h)\), \(i = 1, \ldots, N\).
3. Generate one minibatch of HR face images \(h_i\) from \(x_i\), \(i = 1, \ldots, N\), where \(h_i = \phi(x_i)\).
4. Update the parameters of the discriminative network \(D\).
5. Update the parameters of the generative network \(G\).
6. end while

Outputs: UR-DGN.

This work was supported under the Australian Research Council’s Discovery Projects funding scheme (project DP150104645).
1. MOTIVATION
Our goal is to perform anomaly detection in a unique setting, removing the reliance on data and/or temporal assumptions.

Our setting is largely unaddressed in vision-based anomaly detection, but appears often in practice.

First-time data: New systems and environments
Personalized results: Unique testing distribution
Database sifting: Exploring a single data chunk

Our setting involves two challenging restrictions
(1) Operate relative to the test sequence
(2) Score independent of ordering

2. APPROACH & KEY INSIGHTS
Taking a discriminative, permutation-based approach allows us to operate in this setting

Insight #1: Density ratios directly estimate discriminability, minimizing distribution assumptions

Density ratio concept

How we use density ratio estimation

Insight #2: Permutation testing removes temporal assumptions, avoiding false positives

Scanning techniques
Our method

3. RESULTS
This method performs as well as other methods that require a training set

Avenue Dataset
Similar frame- and pixel-based ROC, without using the training set

Examples: Correct detections
- panda sneeze
- child skipping
- throwing papers
- subtle crowd movement
- single exit-entrance

UMN Dataset
Higher AUC on all but 1 scene
Example: Scene 7

Examples: Failure cases
- crowed running
- illumination
- camera shake

4. SYSTEM OVERVIEW
The framework from video to anomalies

5. FUTURE WORK
Context-driven improvements could come from feature learning, active learning, and data

Feature learning: align with human notion of abnormality
Active learning: incorporating feedback from humans
Datasets: developing larger, more realistic benchmarks

Acknowledgements. This research was supported through the US Department of Defense National Defense Science & Engineering Graduate Fellowship (NDSEG) Program and NSF grant IIS1227495.
Vadim Kantorov, Maxime Oquab, Minsu Cho, Ivan Laptev
INRIA / ENS, Paris

Goal
Localize objects with only image-level labels at train time, i.e. weakly supervised localization.

In train time: ✓ horse ✓ person

Motivation
Weakly supervised localization often results in errors due to:
- shrinking to most discriminative parts
- expansion beyond object boundaries

Weakly supervised localization often results in errors due to:
- shrinking to most discriminative parts
- expansion beyond object boundaries

Key idea
Explore “object contrast” at object boundaries: Maximize the difference of object scores inside and outside object proposals

Contributions
- Explore context-aware ConvNet models for weakly supervised localization (WSL)
- Propose a novel context-aware contrastive model for WSL
- Demonstrate its effectiveness on PASCAL VOC 2007 / 2012
- State of the art on the datasets for models with same base architecture

Task definitions
- Object localization: for a given class label find the most confident tight bounding box
- Object detection: for a given class label find tight bounding boxes for all object instances
- Weakly supervised localization (WSL): localization with only image-level labels at train time
- Evaluated by detection mAP and CorLoc

Related work

Implementation details
- All object proposals with sides greater than 20px are used.
- NVIDIA Titan X single-GPU training
- Optimized by SGD with batch size 1, momentum 0.9 and weight decay 5e-4
- 10 epochs with learning rate 5e-4, then 20 epochs with learning rate 5e-5
- Detections are filtered to have a minimum score of 10^-4, then passed through NMS with threshold 0.4.

Model

<table>
<thead>
<tr>
<th>Model</th>
<th>CorLoc mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>[CVS16]</td>
<td>52.0 30.2</td>
</tr>
<tr>
<td>[WRHT14]</td>
<td>48.5 30.9</td>
</tr>
<tr>
<td>[WRHT14] + context</td>
<td>31.6</td>
</tr>
<tr>
<td>WSDNN-SSW-S [BV16]</td>
<td>31.1</td>
</tr>
<tr>
<td>WSDNN-SSW-ENS [BV16]</td>
<td>54.2 33.3</td>
</tr>
<tr>
<td>WSDNN-SSW-S</td>
<td>50.0 30.5</td>
</tr>
<tr>
<td>additive</td>
<td>52.8 33.3</td>
</tr>
<tr>
<td>contrastive A</td>
<td>50.2 32.2</td>
</tr>
<tr>
<td>contrastive S</td>
<td>55.1 36.3</td>
</tr>
</tbody>
</table>

WSDDN-SSW-S is our reimplementation of [BV16].
Contrastive model S outperforms the baseline [BV16].
About 1% variance for all models measured over 5 runs.

Results on VOC 2012

<table>
<thead>
<tr>
<th>Model</th>
<th>det mAP</th>
<th>corloc mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline [CVS16]</td>
<td>46.4 49.3 47.5 56.9 52.9 54.8</td>
<td></td>
</tr>
<tr>
<td>contrastive S (CorLoc)</td>
<td>54.2 33.3</td>
<td></td>
</tr>
<tr>
<td>contrastive S (mAP)</td>
<td>44.9 49.9 47.5 56.9 52.9 54.8</td>
<td></td>
</tr>
<tr>
<td>contrastive A</td>
<td>50.2 32.2</td>
<td></td>
</tr>
<tr>
<td>contrastive S</td>
<td>55.1 36.3</td>
<td></td>
</tr>
<tr>
<td>baseline (CorLoc)</td>
<td>15.2 27.0 31.1 33.0 30.0 35.3</td>
<td></td>
</tr>
</tbody>
</table>

WSDDN-SSW-S

VOC 2012: 11k trainval images, 11k test images, 20 classes

Example detections
Contrastive S improves over WSDDN-SSW-S

Conclusions
- Contrastive model S finds tight bounding boxes when the baseline [BV16] fails.
- Failure cases include localization of overlapping instances of the same object class.

Torch code is online
http://www.di.ens.fr/willow/research/contextlocnet
Network flow formulations for Learning Binary Hashing

Lopamudra Mukherjee\textsuperscript{1}, Jiming Peng\textsuperscript{2}, Trevor Sigmund\textsuperscript{3}, Vikas Singh\textsuperscript{4}

\textsuperscript{1}University of Wisconsin-Madison, *University of Houston, \textsuperscript{3}University of Wisconsin-Whitewater,

\textsuperscript{1}This work is supported by NSF Career 1252725, NSF CGV 1219016, NIH BD2K 1U54AI117924

INTRODUCTION

Binary Hashing problem
- Map examples in $\mathbb{R}^d$ to binary codes in $\mathbb{B}^d$ where $d \ll D$.
- Numerous applications in high dimensional indexing/retrieval.
- Binary codes are highly storage efficient: only 32 or 64 bits.
- Hamming distance computable efficiently via XOR operations.
- Desiderata: Semantically similar examples should map to similar codes, dissimilar examples map to distant codes.

STARTING POINT OF THIS WORK

Literature and Contributions
- Well studied: Core problem is binary and NP-hard. Generally related problems solved using continuous approaches.
- Investigate the feasibility of simpler discrete energy minimization algorithms.

Can a reformulation enable re-purposing the existing suite of graph-cut based energy minimization solvers?

\textbf{Contribution:} (1) network-flow model for binary hashing; (2) provably partially-optimal solution for each hash code bit.

MAIN IDEAS

Initial Model

Hinge Distance for X's rows

\[
\min_{X} \|A - H\|_F^2 \quad \text{s.t.} \quad H = X(1 - X)^T + (1 - X)X^T, \quad X_{n \times d} \in \{0, 1\}
\]

\[
\text{or} \quad H = -XX^T - XX^T + d_{1 \times n} \cdot \hat{X} = I - X
\]

\[
\min_{\hat{X}} \|\hat{X}X^T + XX^T - A\|_F^2 \quad \text{where} \quad \hat{A} = d_{1 \times n} - A
\]

EXPERIMENTS

Design

- Evaluations on seven benchmarks against nine other approaches.
- Experiments demonstrate scalability and accuracy as a function of number of bits.

\textbf{Objective values on the SSD and} $\ell_1$ \textbf{measures ranked for all methods on toy dataset, with known ground-truth.}

Alg-L1 better than all other algorithms, Alg-SSD is a close second.

Minimization of Objective

\begin{tabular}{l|c|c|c|c}
<table>
<thead>
<tr>
<th>Alg on Obj</th>
<th>Rank 1</th>
<th>Rank 2</th>
<th>Rank 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alg-SSD</td>
<td>97%</td>
<td>3%</td>
<td>0%</td>
</tr>
<tr>
<td>Alg-L1</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Alg-S at SSD</td>
<td>39%</td>
<td>44%</td>
<td>16%</td>
</tr>
<tr>
<td>Alg-E at SSD</td>
<td>43%</td>
<td>45%</td>
<td>11%</td>
</tr>
</tbody>
</table>
\end{tabular}

- Ranking results w.r.t SSD and $\ell_1$ objectives.

\textbf{A network-flow based formulation model for Binary Hashing}

SSD MODEL

Reformulation as Mincut

- Use Sum of Square Distances ($p = 2$) as the distance metric.
- $x, \hat{x}$ are the $i$-th columns of the binary matrices $X$ and $\hat{X}$ respectively. It follows that $x_i + \hat{x}_i = 1_{n \times 1}, x_i^T \hat{x}_i = 0$

\[
\min_{X} \|A - \hat{X}X^T - XX^T\|_F^2
\]

\[
= \min_{X} \|A - \sum_{i=1/Fix} d_{X_i}X^T + \hat{X}_i\|_F^2
\]

Objective $G(x) = \|G - (xx^T + \hat{X}X^T)\|_F^2$ where $x = x_{iter}$. Set $G_0 = G(x) - (G - 1_{n \times n})^2$.

Then $G(x) = \sum_i(G_0 - 1_{n \times n}) + 2 \sum_{i<j} G_i$, where $N_i$ and $N_j$ are two partitions.

Mincut Strategy

1: while convergence not satisfied do
2: for iter = 1 : do
3: Solve Mincut on graph with edge weights $G_0 = G(x) - (G - 1_{n \times n})^2$
4: Update. $x_{iter} = x$
5: end for
6: end while

\textbf{Objective values on the SSD and $\ell_1$ measures ranked for all methods on toy dataset, with known ground-truth.}

Alg-L1 better than all other algorithms, Alg-SSD is a close second.

REGULARIZING FOR BALANCED PARTITIONS

\textbf{Disadvantage:} mincuts may lead to disproportionate cuts.

\textbf{Solution:} Bias the model to produce size regularized cuts:
- Impose a regularizer $R(x) = x_{1 \times n}X^T + (1 - x_{1 \times n})(1 - x)^T$
- $R(x)$ minimized when size of the partitions are same.

\textbf{Objective:} $G(x) + \alpha R(x)$, is quadratic pseudo-boolean function, solved using mincuts, with partially optimal solutions in $\{0, 1\}$

\textbf{How well ‘close’ binary codes represent semantic labels?}

\textbf{Accuracies for Nursery and CIFAR Datasets:} Our approaches find neighbors with the same class label as the query example.
Transfer Neural Trees for Heterogeneous Domain Adaptation

Wei-Yu Chen1,2, Tzu-Ming Hsu2, Yao-Hung Tsai3, Yu-Chiang Frank Wang2, Ming-Syan Chen1
1Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan
2Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
3Department of Machine Learning, Carnegie Mellon University, Pittsburgh, USA

Introduction

- Domain adaptation:
  Address the same learning task across different domains
- Heterogeneous domain adaptation:
  Source and target-domain data are described by distinct types of features.

Related Works

- Map cross-domain data onto a common subspace for classification
  ✓ Jointly learning of mapping & classification functions from cross-domain labeled data [1]
  ✓ Project cross-domain data for suppressing domain differences [2]
  ✓ A pair of DNNs with shared parameters for matching cross-domain data. [3]

Proposed Method

Transfer Neural Trees (TNT)
- Source-domain mapping \( f_S \)
- Target-domain mapping \( f_T \)
- Prediction layer \( G \)

- Learning \( f_S \) and \( G \):
  ✓ Minimize prediction loss \( L_p \) of source-domain data \( X_S \)
  \[
  \min_{f_S, G} \sum_{(x,y) \in D_L} L_p(\theta, x, y) \quad \text{where} \quad L_p(\theta, x, y) = \sum_{z \in D_L} \log(\tilde{y}(x|\theta, z))
  \]

- Learning \( f_T \) with fixed \( G \) (semi-supervised learning):
  ✓ Minimize prediction loss \( L_p \) of target-domain labeled data \( X_L \)
  ✓ Minimize embedding loss \( L_e \) of target-domain labeled & unlabeled data
  \[
  \min_{f_T} \sum_{(x,y) \in D_L} L_p(\theta, x, y) + \lambda \sum_{(x,y) \in D_U} L_e(\theta, x, y)
  \]
  where \( L_p(\theta, x, y) = \sum_{z \in D_L} \sum_{(x,y) \in D_L} L_p(\theta, x, y, z) \)
  \( L_e(\theta, x, y) = \sum_{z \in D_U} L_p(\theta, x, y, z) \)

  ✓ Increase prediction consistency \( \Rightarrow \) preserve structural consistency btw \( X_L \) & \( X_U \)

Experiments

- Datasets
  ✓ Object recognition (10 classes):
    - Amazon: 959 DeCAF/SURF features
    - Webcam: 296 DeCAF/SURF features
    - Caltech: 1124 DeCAF/SURF features
  ✓ Text-to-image recognition (8 classes):
    - NUS-WIDE tag data: 800 NN features
    - ImageNet image: 800 DeCAF features

- Settings
  ✓ Source domain: all data in dataset as \( X_S \)
  ✓ Target domain: 3 per class as \( X_A \), the rest as \( X_U \)

Evaluation

- Cross Features
  ✓ Cross Datasets
  ✓ Cross Modalities

Visualization

- Different \( G \) in TNT

Conclusions

- TNT for semi-supervised & cross-domain deep learning
- Transfer-NDF with stochastic pruning for HDA
- Embedding loss in TNT for preserving prediction & structural consistency
- Promising results on cross-feature, domain, and modality classification tasks

Reference

Tracking Persons-of-Interest via Adaptive Discriminative Features

Shun Zhang1 Yihong Gong1 Jia-Bin Huang2 Jongwoo Lim3 Jinjun Wang1 Narendra Ahuja2 Ming-Hsuan Yang4
1Xi’an Jiaotong University 2University of Illinois, Urbana-Champaign 3Hanyang University 4University of California, Merced


Problem

- Frequent shot changes
- Large face appearance variations
- Low resolution, motion blurring, occlusion and so on

Contributions

- Learning video-specific features by adapting deep CNN based on contrastive and triplet loss
- An improved symmetric triplet loss function (SymTriplet)
- A fully automatic multi-face tracking algorithm (detection, tracking, clustering, and feature adaptation)
- A new dataset with 8 music videos from YouTube

Overview

(a) Pre-training
(b) Automatic training sample discovery
(c) Adaptive feature learning
(d) Linking tracklets

Adaptive Feature Learning

Siamese Network

Contrastive Loss

Triplet Network

SymTriplet loss:

\[ L_{\alpha} = \max \left\{ 0, d(x_i, x_p) - \frac{1}{2} d(x_i, x_n) - d(x_p, x_n) + \alpha \right\} \]

Weighted Purity vs. #Clusters

Quantitative Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall↑</th>
<th>Precision↑</th>
<th>F1↑</th>
<th>FAF↓</th>
<th>IDS↓</th>
<th>Frag↓</th>
<th>MOTA↑</th>
<th>MOTP↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>mTLD [Kalal et al. 2012]</td>
<td>9.7%</td>
<td>36.1%</td>
<td>15.3%</td>
<td>0.39</td>
<td>280</td>
<td>621</td>
<td>68.4%</td>
<td></td>
</tr>
<tr>
<td>ADMM [Ayazoglu et al. 2012]</td>
<td>75.5%</td>
<td>61.8%</td>
<td>68.0%</td>
<td>0.50</td>
<td>2382</td>
<td>2959</td>
<td>51.7%</td>
<td>63.7%</td>
</tr>
<tr>
<td>IHTLS [Dicle et al. 2013]</td>
<td>75.5%</td>
<td>68.0%</td>
<td>71.6%</td>
<td>0.41</td>
<td>2013</td>
<td>2880</td>
<td>56.2%</td>
<td>63.7%</td>
</tr>
<tr>
<td>Pre-trained</td>
<td>60.1%</td>
<td>88.8%</td>
<td>71.7%</td>
<td>0.17</td>
<td>931</td>
<td>2140</td>
<td>51.5%</td>
<td>79.5%</td>
</tr>
<tr>
<td>Ours-Siamese</td>
<td>71.5%</td>
<td>89.4%</td>
<td>79.4%</td>
<td>0.19</td>
<td>986</td>
<td>2512</td>
<td>62.3%</td>
<td>64.0%</td>
</tr>
<tr>
<td>Ours-Triplet</td>
<td>71.8%</td>
<td>88.8%</td>
<td>79.4%</td>
<td>0.20</td>
<td>986</td>
<td>2546</td>
<td>61.8%</td>
<td>64.2%</td>
</tr>
<tr>
<td>Ours-SymTriplet</td>
<td>71.8%</td>
<td>89.7%</td>
<td>79.8%</td>
<td>0.19</td>
<td>986</td>
<td>2563</td>
<td>62.8%</td>
<td>64.3%</td>
</tr>
</tbody>
</table>

Qualitative Results
EXTENDING LONG SHORT-TERM MEMORY FOR MULTI-VIEW STRUCTURED LEARNING

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis and Roland Goecke

Sshyam.Rajagopalan@canberra.edu.au  morency@cs.cmu.edu  tbaltrus@cs.cmu.edu  roland.goecke@ieee.org

Multi-View Learning

1. View – a particular way of observing a phenomena. Eg. (a) Image and text for image captioning, (b) Headpose, HOG, HOF for videos

2. Multi-View Interactions – View-specific dynamics capture interactions between hidden outputs from the same view, while cross-view dynamics capture interactions between hidden outputs of different views. Both interactions are very common in many problems.

Multi-View Learning

Deep Multi-View Representation Learning Models - CCA + Autoencoder, LSTM as language decoder integrating images and text, Multimodal LSTM for speaker identification. Models are applied to behaviour recognition and image captioning problems.

Existing models lack flexibility in designing multiple topologies to model view-specific and cross-view interactions.

Model Analysis

Experiment Results – Child’s Engagement Level Prediction – Multimodal Dyadic Behavior Dataset

<table>
<thead>
<tr>
<th>Class labels</th>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy to engage</td>
<td>LSTM (Early fusion)</td>
<td>0.75</td>
<td>0.81</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM Full</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM Coupled</td>
<td>0.79</td>
<td>0.81</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM Hybrid</td>
<td>0.80</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>Difficult to engage</td>
<td>LSTM (Early fusion)</td>
<td>0.63</td>
<td>0.55</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM Full</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM Coupled</td>
<td>0.67</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM Hybrid</td>
<td>0.74</td>
<td>0.64</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Experiment Results – Image Caption Generation

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flickr8K</td>
<td>Log Bilinear</td>
<td>65.6</td>
<td>42.4</td>
<td>27.7</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>NIC</td>
<td>63.0</td>
<td>41.0</td>
<td>27.0</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>BRNN</td>
<td>57.9</td>
<td>38.3</td>
<td>24.5</td>
<td>19.5</td>
</tr>
<tr>
<td></td>
<td>Soft Attention</td>
<td>67.0</td>
<td>44.8</td>
<td>29.9</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>Hard Attention</td>
<td>67.0</td>
<td>45.7</td>
<td>31.4</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>gLSTM</td>
<td>64.7</td>
<td>45.9</td>
<td>31.8</td>
<td>21.6</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM</td>
<td>65.7</td>
<td>46.9</td>
<td>32.6</td>
<td>22.2</td>
</tr>
<tr>
<td>Flickr30K</td>
<td>Log Bilinear</td>
<td>60.0</td>
<td>38.0</td>
<td>25.4</td>
<td>17.1</td>
</tr>
<tr>
<td></td>
<td>NIC</td>
<td>66.3</td>
<td>42.3</td>
<td>27.7</td>
<td>18.3</td>
</tr>
<tr>
<td></td>
<td>BRNN</td>
<td>57.3</td>
<td>36.9</td>
<td>24.0</td>
<td>15.7</td>
</tr>
<tr>
<td></td>
<td>Soft Attention</td>
<td>66.7</td>
<td>43.4</td>
<td>28.8</td>
<td>19.1</td>
</tr>
<tr>
<td></td>
<td>Hard Attention</td>
<td>66.9</td>
<td>43.9</td>
<td>29.6</td>
<td>19.9</td>
</tr>
<tr>
<td></td>
<td>gLSTM</td>
<td>64.6</td>
<td>44.6</td>
<td>30.5</td>
<td>20.6</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM</td>
<td>64.5</td>
<td>44.6</td>
<td>31.1</td>
<td>21.2</td>
</tr>
<tr>
<td>MS-COCO</td>
<td>Log Bilinear</td>
<td>70.8</td>
<td>48.9</td>
<td>34.4</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td>NIC</td>
<td>66.6</td>
<td>46.1</td>
<td>32.9</td>
<td>24.6</td>
</tr>
<tr>
<td></td>
<td>BRNN</td>
<td>62.5</td>
<td>45.0</td>
<td>32.1</td>
<td>23.0</td>
</tr>
<tr>
<td></td>
<td>Soft Attention</td>
<td>70.7</td>
<td>49.2</td>
<td>34.4</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td>Hard Attention</td>
<td>71.8</td>
<td>50.4</td>
<td>35.7</td>
<td>25.0</td>
</tr>
<tr>
<td></td>
<td>gLSTM</td>
<td>67.0</td>
<td>49.1</td>
<td>35.8</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td>MV-LSTM</td>
<td>69.1</td>
<td>51.5</td>
<td>37.7</td>
<td>27.6</td>
</tr>
</tbody>
</table>

Conclusion

1. Extended LSTM for designing multiple topologies to model view relationships
2. Cross-view learning using the proposed model helps in better representation for behaviour recognition and image captioning
1. Main Contributions

- An undirected complete labeled graph for action representation. A node of the graph is modeled to a motionlet which is a semantic part of the trajectory of a joint. The edge is labeled by spatio-temporal relationships between the connected motionlets. A subgraph of this graph can be viewed as a discriminative subaction and has rich semantic information.

- A novel graph kernel called SPGK. Graphs are decomposed into several subgraphs and subgraph matching corresponds to part-based comparison between subactions. Since these substructures capture rich topological structure of graph compared to nodes, it is more discriminative and robust than local or global feature-based comparison.

2. Graph Based Action Representation

- The extracted motionlets by preprocessing raw trajectories of joints and segmenting the processed trajectories

- Constructing a graph composed of obtained motionlets for a video action

3. Subgraph-Pattern Graph Kernel for Action Recognition

- Two basic kernels on nodes and edges of graphs

$$k_v(v_i, v_p) = \begin{cases} 1 & \text{if } \alpha(v_i) = \alpha(v_p) \\ \gamma & \text{if } (\alpha(v_i), \alpha(v_p)) \in S \\ 0 & \text{otherwise,} \end{cases}$$

$$k_e(e_{ij}, e_{pq}) = \frac{1}{2}[k_e(e_{ij}, e_{pq}) + k_e(e_{ij}, e_{pq})]$$

4. Experiments

Results on the MSR Action3D Dataset

5. Conclusions

- We proposed an undirected complete labeled graph to model human actions.

- A novel subgraph-pattern graph kernel (SPGK) was proposed to measure the similarity between two videos by combining all comparisons between subgraphs extracted from two videos.

- We evaluated our algorithm through a series of experiments and demonstrated that our method achieves a comparable performance.
**Motivation**

Both photometric and geometric properties of the scene are used for the confidence estimation.

- $P_{AI}$: amplitude and intensity of ToF signal.
  \[ \sigma_z = \frac{c}{4\pi f_{mod}} \sqrt{\frac{I}{A}} \Rightarrow \sigma_d = b f \sigma_z \zeta^2 - \sigma_z^2 \]  
- $P_{LV}$: accounts for local depth variance.
  \[ D_{ToF}^2 = \frac{1}{|N(p_{ToF})|} \sum_{j \in N(p_{ToF})} |z_i - z_j| \]

**Data Fusion**

Enforce local consistency and weight the two contributions exploiting the confidence information.

\[ \Omega_f(d) = \sum_{g \in A} \left( P_f(g) P_{f,g,T}(d) + P_s(g) P_{f,g,S}(d) \right) \]

where $P_{f,g}(d) = e^{-\frac{D_{ToF}^2}{\gamma}} \cdot e^{-\frac{\Delta f}{\gamma}} \cdot e^{-\frac{\Delta f'}{\gamma}}$, $f, g$ and $f', g'$: points in the left and right image.

- $\Delta$: spatial proximity; $\Delta^v, \Delta^c$: color similarity

**Proposed Method**

- ToF Data
- Interpretation of ToF Data
- ToF Confidence
- Stereo Disparity Map and Confidence
- Locally Consistent Fusion
- Output Disparity Map

**Experimental Results - Comparison**

<table>
<thead>
<tr>
<th>Scene</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToF Int.</td>
<td>9.83</td>
<td>10.33</td>
<td>14.43</td>
<td>8.68</td>
<td>15.12</td>
<td>11.67</td>
</tr>
<tr>
<td>Stereo</td>
<td>19.17</td>
<td>27.83</td>
<td>18.06</td>
<td>25.52</td>
<td>11.49</td>
<td>20.42</td>
</tr>
<tr>
<td>Fusion</td>
<td>7.40</td>
<td>9.33</td>
<td>6.92</td>
<td>6.30</td>
<td>8.39</td>
<td>7.67</td>
</tr>
</tbody>
</table>

**Stereo Disparity Map and Confidence**

- A high resolution disparity map is inferred by global or semi-global stereo vision algorithms.
- Confidence
  \[ P_s = \Delta C^d \left( 1 - \frac{\min(\Delta d^s, \gamma)}{\gamma} \right) \]
  \[ \frac{C_1}{C_1} = \frac{C_1}{C_1} \]
- Local cost, $\gamma$: Global cost
- $\Delta C^d = C^d - C^d$, $\Delta d^s = |d^s - d^s|$
### Fast, Exact & Multi-Scale Inference for Semantic Image Segmentation with Deep Gaussian CRFs

Siddhartha Chandra & Iasonas Kokkinos

INRIA GALEN & Centrale Supélec Paris

#### Gaussian Conditional Random Fields

![Gaussian CRF Diagram]

Unique and exact global optimum, pairwise interactions discovered from the data via end-to-end deep learning, and fast inference via efficient implementation.

#### Quadratic Energy Optimization

\[ E(x) = \frac{1}{2} x^T(A + M)x - Bx \]  

If \((A + M)\) is symmetric positive definite, unique global minimum at:

\[ (A + M)x = B. \]  

Inference involves solving a system of linear equations.

#### Quadratic Optimization in Deep Learning

- Network populates unary and pairwise terms
- \(QO\) module proposes scores after inference
- Model parameters learnt end-to-end for arbitrary global loss (objective) \(L\)
- Gradient expressions:
  - Using Cholesky decomposition, equality: \(\sum_{ij} A_{ij} x_i x_j = b_i\)  
- Gradient computed analytically by solving a system of linear equations.

#### Potts Type Model with Shared Pairwise Terms

**Notation:** \(A_{\alpha,\beta}((l_i, l_j))\) is the pairwise energy term for pixels \(p_i, p_j\), taking the labels \(l_i\) and \(l_j\). Per-class scores and unaries are denoted by \(x_i\) and \(b_i\), where \(k \in \{1, \ldots, |L|\}\).

\[ A_{\alpha,\beta}((l_i, l_j)) = \begin{cases} 0 & l_i = l_j \\ A_{\alpha,\beta} & l_i \neq l_j \end{cases} \]  

- Fewer parameters \((P \times P)\) compared to general setting \((PL \times PL)\) terms for \(P\) pixels, \(L\) labels. Reduction factor of \(441\) for VOC Pascal.
- Algebraic simplifications enable us to infer scores for each label independently

\[ \sum_{i=1}^{L} x_i = \sum_{i=1}^{P} b_i. \]  

- Training \(3x\) faster, inference \(6x\) faster than general setting on VOC Pascal

#### Implementation Details and Efficiency

- Conjugate Gradient > other algorithms
- CUDA-based implementation using efficient CUDA Sparse, Blue routines
- General Inference time \(\sim 0.02s\)
- Potts-type inference time \(\sim 0.003s\)
- Code available at [https://github.com/siddharthachandra/gcrf](https://github.com/siddharthachandra/gcrf)

#### Experimental Setup

- All methods use VOC PASCAL 2012 image segmentation benchmark
- Basenet is a 3-resolution variant of Deeplab-LargeFOV
- We experiment with 4 variants of our method
  - \(QO\) General pairwise terms
  - \(QO^P\) - Potts-shared pairwise terms
  - \(QO^{P+}\): One \(QO\) per resolution
  - \(QO^{P++}\): Multi-resolution \(QO\)

#### Acknowledgements

This work has been funded by the EU Projects MOBOT FP7-ICT-2011-600796 and I-SUPPORT 643666 #2020.
Kernel-Based Supervised Discrete Hashing for Image Retrieval
Xiaohuang Shi, Fuyong Xing, Jinzheng Cai, Zizhao Zhang, Yuanpu Xie and Lin Yang
University of Florida

Abstract
In this paper, we propose a novel yet simple kernel-based supervised discrete hashing method via an asymmetric relaxation strategy. Specifically, we present an optimization model with preserving the hashing function and the relaxed linear function simultaneously to reduce the accumulated quantization error between hashing and linear functions. Furthermore, we improve the hashing model by relaxing the hashing function into a general binary code matrix and introducing an additional regularization term. Then we solve these two optimization models via an alternative strategy, which can effectively and stably preserve the similarity of neighbors in a low-dimensional Hamming space. The proposed hashing method can produce informative short binary codes that require less storage volume and lower optimization time cost. Extensive experiments on multiple benchmark databases demonstrate the effectiveness of the proposed hashing method with short binary codes and its superior performance over the state of the arts.

Objective

Kernel supervised discrete hashing with hashing function preserved (KSDH_H)
Objective function
\[
\min_{A} \| H^T A K - r S \|_F^2 \\
\text{s.t. } A K^T A^T = n I, H = sgn(AK)
\]
The hashing function \( H \) is preserved in the objective function. The constraint \( A K^T A^T = n I \), is derived from the constraint \( H H^T = n I \), which enforces \( r \) bit hashing codes mutually uncorrelated such that the redundancy among these bits is minimized.

Kernel supervised discrete hashing with a relaxed binary code matrix (KSDH_B)
Objective function
\[
\min_{A} \| B^T A K - r S \|_F^2 + \lambda \| B - A K \|_F^2 \\
\text{s.t. } A K^T A^T = n I
\]
\( B \) represents the binary codes of training data, the term \( \| B - A K \|_F^2 \) aims to reduce the accumulated errors, \( \lambda \) is to balance the semantic information and accumulated errors. In addition, the regularization term can guarantee the objective function to have a stable optimal solution.

Motivations

- NP-hard objective function—there is no exact solution known
- Symmetric relaxation—relaxing the discrete matrices into continuous matrices can make the problem easily be solved, while it would generate accumulated quantization errors between the discrete and continuous matrices
- Discrete matrix preservation—directly learning binary codes can reduce the accumulated errors.

NP-hard objective
\[
\min_{A} \| sgn(AK)^T sgn(AK) - r S \|_F^2
\]
Symmetric relaxation
\[
\min_{A} \| (A K)^T (A K) - r S \|_F^2
\]

We evaluate KSDH_H and KSDH_B on four publicly available benchmark databases: CIFAR-10, MNIST, Youtube and ImageNet. We compare the proposed algorithms against SSH [1], BRE [2], KSH [3] and SDH [4]. For the CIFAR-10 database, we partition it into two parts: a training subset of 59K images and a test query set 1K images, which contains ten categories with each consisting of 100 images. We uniformly select 100 and 500 images from each category to form two training sets, respectively.

Experiment
- We evaluate KSDH_H and KSDH_B on four publicly available benchmark databases: CIFAR-10, MNIST, Youtube and ImageNet. We compare the proposed algorithms against SSH [1], BRE [2], KSH [3] and SDH [4]. For the CIFAR-10 database, we partition it into two parts: a training subset of 59K images and a test query set 1K images, which contains ten categories with each consisting of 100 images. We uniformly select 100 and 500 images from each category to form two training sets, respectively.

Reference:
Iterative Reference Driven Metric Learning for Signer Independent Isolated Sign Language Recognition

Fang Yin, Xiujuan Chai, Xilin Chen
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
{fang.yin, xiujuan.chai, xilin.chen}@vipl.ict.ac.cn

Motivation

Sign Language Recognition (SLR) aims to bridge the gap between the hard of hearing and the hearing, and alleviate the social isolation for the hard of hearing.

Motivation

- There exists big variations among different signers, due to the different heights of the signers or the habits during signing.
- A realistic SLR system must overcome the challenges brought by the inter-signer variations.

The Proposed Method

Basic Idea

- Use a signer invariant reference to represent each sign, and learn the metric accordingly.
- Optimize the reference and the metric alternately.

Framework of iRDML

- RDML is proposed to learn the distance between specific references and the training samples.
- iRDML is designed based on RDML to further explore more appropriate references and the corresponding distance metric.
- The effectiveness and efficiency of the proposed method is evaluated extensively on several public databases for both sign language recognition and human motion recognition tasks.

Evaluation on DEVISIGN Dataset

- Dataset
  DEVISIGN (Chai, TR’14) dataset has a vocabulary size of 2000 from 8 different signers.
  Evaluation protocol: 8 groups of data from 4 signers form the training set and the data from other 4 signers are the test data.
  URL: vipl.ict.ac.cn/homepage/KSL/data.html

- Comparison between RDML and iRDML

- Comparison with existing methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Chai et al. (Chai, TR’14)</th>
<th>LMNN</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMM</td>
<td>57.4</td>
<td>55.9</td>
<td>57.1</td>
</tr>
<tr>
<td>DTW</td>
<td>61.7</td>
<td>66.6</td>
<td>65.8</td>
</tr>
<tr>
<td>ARMA</td>
<td>65.8</td>
<td>64.9</td>
<td>67.1</td>
</tr>
<tr>
<td>ITML</td>
<td>68.0</td>
<td>67.8</td>
<td>67.1</td>
</tr>
<tr>
<td>CSML</td>
<td>70.5</td>
<td>73.4</td>
<td>71.7</td>
</tr>
<tr>
<td>LMNN/iRDML</td>
<td>70.0</td>
<td>72.1</td>
<td>71.7</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td>66.8</td>
</tr>
</tbody>
</table>

Evaluation on Human Motion Recognition

- Dataset
  HDM05 (Cho, VISAPP’14) consists of 2337 motion sequences from 65 actions.
  Evaluation protocol: 10-fold cross validation.

- Comparison with existing methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Cho et al. (Cho, VISAPP’14)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELM</td>
<td>91.57</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>94.95</td>
<td></td>
</tr>
<tr>
<td>MLP</td>
<td>95.20</td>
<td></td>
</tr>
<tr>
<td>Hybrid MLP</td>
<td>95.59</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>95.76</td>
</tr>
</tbody>
</table>

Conclusion

- RDML is proposed to learn the distance between specific references and the training samples.
- iRDML is designed based on RDML to further explore more appropriate references and the corresponding distance metric.
- The effectiveness and efficiency of the proposed method is evaluated extensively on several public databases for both sign language recognition and human motion recognition tasks.
**GOAL**

- Task
  - Visual Question Answering (VQA)
  - Answer a question about a given photograph
- Applications
  - Assist the visually impaired
  - Automatically query surveillance video

**CONTRIBUTIONS**

- Existing Methods
  - End-to-end deep VQA networks adapted from captioning models: utilize a recurrent LSTM network, which takes the question and CNN image features as input and outputs the answer. [6, 7]
- Problems
  - Do not have any explicit notion of object position
  - Use the whole question encoding to infer the answer, without considering fine-grained information from the question
- Contributions
  - Propose Spatial Memory Network VQA (SMem-VQA)
  - Incorporate explicit spatial attention based on memory networks
  - Use fine-grained word embeddings to collect visual evidence for each word in the question

**SCHEMATIC DIAGRAM**

Attention is applied in two steps (hops):

1. **is the child standing on?**
2. **stateboard**

**SYNTHETIC EXPERIMENTS**

By visualizing attention, we can figure out how the network learns to answer questions.

- Absolute Position Recognition
  - Input image: a red square appears in one of the four regions of a white-background image
  - Question: Is there a red square on the [top/bottom/left/right]?

- Network learned two logic rules
  - Look at the position specified in question [top/bottom/right], if it contains a square, then answer “yes”, if not, then answer “no”.
  - Look at the region where there is a square, then answer “yes” for the question about that position and “no” for the questions about the other three positions.

- Relative Position Recognition

**SMem-VQA NETWORK ARCHITECTURE**

![SMem-VQA Network Architecture Diagram](image)

**EXPERIMENTAL RESULTS**

Test-dev and test-standard results on Open-Ended VQA dataset [1] (accuracy). Models with = use extra training data in addition to the VQA dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Test-dev</th>
<th>Test-standard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>overall</td>
<td>yes/no</td>
</tr>
<tr>
<td>LSTM CH [1]</td>
<td>59.74</td>
<td>79.94</td>
</tr>
<tr>
<td>AKIQ [2]</td>
<td>58.72</td>
<td>79.23</td>
</tr>
<tr>
<td>SOPW [2]</td>
<td>57.23</td>
<td>80.71</td>
</tr>
<tr>
<td>SMem-VQA</td>
<td>56.56</td>
<td>75.98</td>
</tr>
</tbody>
</table>

- 0.1 accuracy result on the reduced DAQUAR dataset [5] is 40.07%.
- Per-answer category attention weight visualization analysis:

  - year/no: (ic3) the(2) cow(27) young(38) ? yes others: where(26) are(5) the(2) cakes(16) ? on table

  - number: how(10) many(11) striped(902) pillows(405) are(5) in(7) the(2) sofa(309) ? 4

**REFERENCES**

We proposed the method **Relay Backpropagation**, which encourages the flows of informative gradient in backward propagation when training deep convolutional neural networks. Relevant information can be effectively preserved, and the adverse effect of less relevant information can be restrained.

We achieved the **1st place in ILSVRC 2015 Scene Classification Challenge**. Extensive experiments on two large scale challenging datasets demonstrate the effectiveness of our method is not restricted to a specific dataset or network architecture.

**Relay Backpropagation**

<table>
<thead>
<tr>
<th>Standard BP</th>
<th>Relay BP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Diagram" /></td>
<td><img src="image2" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**Results on Places2**

Single crop error rates (%) on Places2 challenge validation set. In the brackets are the improvements over "standard BP" baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>top-1 err.</th>
<th>top-5 err.</th>
<th>top-1 err.</th>
<th>top-5 err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard BP</td>
<td>50.91</td>
<td>19.00</td>
<td>50.62</td>
<td>19.69</td>
</tr>
<tr>
<td>multi-loss + standard BP</td>
<td>50.72(0.19)</td>
<td>18.84(0.18)</td>
<td>50.56(0.03)</td>
<td>18.65(0.61)</td>
</tr>
<tr>
<td>Relay BP</td>
<td>49.75(1.16)</td>
<td>17.86(0.17)</td>
<td>49.77(0.55)</td>
<td>17.89(0.43)</td>
</tr>
</tbody>
</table>

Error rates (%) on Places2 test set [3]. Our team “WM” won the 1st place. Our single-model results outperform all ensemble results of other teams.

**Insight**

**Difficulty**: The improvement on accuracy cannot trivially achieved by simply increasing the depth of network.


**Our insight**: Although the gradient does not vanish, less relevant information derived from loss is received by lower layers. Hence for effective update of the weights, BP should not go back too many layers.

**Network Architectures for Places2**

<table>
<thead>
<tr>
<th>input size</th>
<th>gradient</th>
<th>model A</th>
<th>model B</th>
</tr>
</thead>
<tbody>
<tr>
<td>224×224</td>
<td><img src="image3" alt="Diagram" /></td>
<td><img src="image4" alt="Diagram" /></td>
<td></td>
</tr>
<tr>
<td>112×112</td>
<td><img src="image5" alt="Diagram" /></td>
<td><img src="image6" alt="Diagram" /></td>
<td></td>
</tr>
<tr>
<td>56×56</td>
<td><img src="image7" alt="Diagram" /></td>
<td><img src="image8" alt="Diagram" /></td>
<td></td>
</tr>
<tr>
<td>28×28</td>
<td><img src="image9" alt="Diagram" /></td>
<td><img src="image10" alt="Diagram" /></td>
<td></td>
</tr>
<tr>
<td>14×14</td>
<td><img src="image11" alt="Diagram" /></td>
<td><img src="image12" alt="Diagram" /></td>
<td></td>
</tr>
</tbody>
</table>

**Results on ImageNet 2012**

**Single model error rates (%) on ImageNet 2012 classification dataset.**

<table>
<thead>
<tr>
<th>Method</th>
<th>dataset</th>
<th>ResNet-50</th>
<th>Inception-v3</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard BP</td>
<td><img src="image13" alt="Diagram" /></td>
<td>20.74</td>
<td>5.25</td>
</tr>
<tr>
<td>Relay BP</td>
<td><img src="image14" alt="Diagram" /></td>
<td>21.17</td>
<td>5.37</td>
</tr>
<tr>
<td>Relay BP (re-implement)</td>
<td><img src="image15" alt="Diagram" /></td>
<td>20.26</td>
<td>4.93</td>
</tr>
</tbody>
</table>

References


Counting in the Wild
Carlos Arteta¹, Victor Lempitsky² and Andrew Zisserman¹
¹University of Oxford, UK ²Skolkovo Institute of Science and Technology (Skoltech), Russia

Objectives and Contributions
- Learning to count from crowdsourced dot-annotations.
- Application: penguins in the wild.
- Challenge: weak and noisy annotations (e.g. systematic undercounting).
- Challenge: complex background and penguin appearance variability.
- We propose a multi-task network for counting, segmenting and predicting annotators’ disagreement from crowdsourced dot-annotations.
- We release the first version of the large penguin dataset.

The Penguin Dataset
- Ongoing project to monitor penguin colonies in Antarctica.
- Already available[1]: Over 80k large images from 36 fixed cameras.
- Annotations are provided by citizen scientists who volunteer to use the tool [2]; the current release of the dataset is annotated by an average of 8.75 users per image.
- Two splits of the data are provided: one with mixed cameras and one where cameras are kept separate.

Model
- Learn mapping from images to pixel-wise object density, as well as predict the difficulty of the counting task across the image.
- Multi-task CNN which leverages segmentation to define the labels for density regression; even rough a segmentation provides spatial cues and correspondence between dots and groups of objects.
- Crowdsourced labels provide information about the difficulty of the task through the agreement between annotators.

Learning with Weak/Noisy Supervision
- How can we define the density regression labels from dot-annotations?
  The object density regression builds on top of a foreground/background segmentation which can be easier to learn from crowdsourced dot-annotations.
- How can we define the labels for binary foreground-background segmentation from dot-annotations?
  We define the segmentation labels as trimaps. These can be built using penguin size information (depth) or by exploiting the dispersion in the dot-annotations of different users (without depth).

Learning with Weak/Noisy Supervision

- How can we define the density regression labels from dot-annotations?
  The density regression builds on top of a foreground/background segmentation which can be easier to learn from crowdsourced dot-annotations.

- How can we define the labels for binary foreground-background segmentation from dot-annotations?
  We define the segmentation labels as trimaps. These can be built using penguin size information (depth) or by exploiting the dispersion in the dot-annotations of different users (without depth).

References
[1] Penguin research webpage, www.robots.ox.ac.uk/~vgg/research/penguins

Results

- Ongoing project to monitor penguin colonies in Antarctica.
- Already available[1]: Over 80k large images from 36 fixed cameras.
- Annotations are provided by citizen scientists who volunteer to use the tool [2]; the current release of the dataset is annotated by an average of 8.75 users per image.
- Two splits of the data are provided: one with mixed cameras and one where cameras are kept separate.

- We propose a multi-task network for counting, segmenting and predicting annotators’ disagreement from crowdsourced dot-annotations.

- We release the first version of the large penguin dataset.

Examples
Crowdsourced dot-annotations

Dot-annotations by 8 different volunteer annotators (colour-coded).

Model

- Learn mapping from images to pixel-wise object density, as well as predict the difficulty of the counting task across the image.
- Multi-task CNN which leverages segmentation to define the labels for density regression; even rough a segmentation provides spatial cues and correspondence between dots and groups of objects.
- Crowdsourced labels provide information about the difficulty of the task through the agreement between annotators.

Learning with Weak/Noisy Supervision

- How can we define the density regression labels from dot-annotations?
  The object density regression builds on top of a foreground/background segmentation which can be easier to learn from crowdsourced dot-annotations.

- How can we define the labels for binary foreground-background segmentation from dot-annotations?
  We define the segmentation labels as trimaps. These can be built using penguin size information (depth) or by exploiting the dispersion in the dot-annotations of different users (without depth).

References
[1] Penguin research webpage, www.robots.ox.ac.uk/~vgg/research/penguins

Results

- Ongoing project to monitor penguin colonies in Antarctica.
- Already available[1]: Over 80k large images from 36 fixed cameras.
- Annotations are provided by citizen scientists who volunteer to use the tool [2]; the current release of the dataset is annotated by an average of 8.75 users per image.
- Two splits of the data are provided: one with mixed cameras and one where cameras are kept separate.

- We propose a multi-task network for counting, segmenting and predicting annotators’ disagreement from crowdsourced dot-annotations.

- We release the first version of the large penguin dataset.

Examples
Crowdsourced dot-annotations

Dot-annotations by 8 different volunteer annotators (colour-coded).

Model

- Learn mapping from images to pixel-wise object density, as well as predict the difficulty of the counting task across the image.
- Multi-task CNN which leverages segmentation to define the labels for density regression; even rough a segmentation provides spatial cues and correspondence between dots and groups of objects.
- Crowdsourced labels provide information about the difficulty of the task through the agreement between annotators.

Learning with Weak/Noisy Supervision

- How can we define the density regression labels from dot-annotations?
  The object density regression builds on top of a foreground/background segmentation which can be easier to learn from crowdsourced dot-annotations.

- How can we define the labels for binary foreground-background segmentation from dot-annotations?
  We define the segmentation labels as trimaps. These can be built using penguin size information (depth) or by exploiting the dispersion in the dot-annotations of different users (without depth).

References
[1] Penguin research webpage, www.robots.ox.ac.uk/~vgg/research/penguins

Results

- Ongoing project to monitor penguin colonies in Antarctica.
- Already available[1]: Over 80k large images from 36 fixed cameras.
- Annotations are provided by citizen scientists who volunteer to use the tool [2]; the current release of the dataset is annotated by an average of 8.75 users per image.
- Two splits of the data are provided: one with mixed cameras and one where cameras are kept separate.

- We propose a multi-task network for counting, segmenting and predicting annotators’ disagreement from crowdsourced dot-annotations.

- We release the first version of the large penguin dataset.

Examples

Crowdsourced dot-annotations

Dot-annotations by 8 different volunteer annotators (colour-coded).

Model

- Learn mapping from images to pixel-wise object density, as well as predict the difficulty of the counting task across the image.
- Multi-task CNN which leverages segmentation to define the labels for density regression; even rough a segmentation provides spatial cues and correspondence between dots and groups of objects.
- Crowdsourced labels provide information about the difficulty of the task through the agreement between annotators.

Learning with Weak/Noisy Supervision

- How can we define the density regression labels from dot-annotations?
  The density regression builds on top of a foreground/background segmentation which can be easier to learn from crowdsourced dot-annotations.

- How can we define the labels for binary foreground-background segmentation from dot-annotations?
  We define the segmentation labels as trimaps. These can be built using penguin size information (depth) or by exploiting the dispersion in the dot-annotations of different users (without depth).

References
[1] Penguin research webpage, www.robots.ox.ac.uk/~vgg/research/penguins

Results

- Ongoing project to monitor penguin colonies in Antarctica.
- Already available[1]: Over 80k large images from 36 fixed cameras.
- Annotations are provided by citizen scientists who volunteer to use the tool [2]; the current release of the dataset is annotated by an average of 8.75 users per image.
- Two splits of the data are provided: one with mixed cameras and one where cameras are kept separate.

- We propose a multi-task network for counting, segmenting and predicting annotators’ disagreement from crowdsourced dot-annotations.

- We release the first version of the large penguin dataset.

Examples

Crowdsourced dot-annotations

Dot-annotations by 8 different volunteer annotators (colour-coded).

Model

- Learn mapping from images to pixel-wise object density, as well as predict the difficulty of the counting task across the image.
- Multi-task CNN which leverages segmentation to define the labels for density regression; even rough a segmentation provides spatial cues and correspondence between dots and groups of objects.
- Crowdsourced labels provide information about the difficulty of the task through the agreement between annotators.
A Discriminative Feature Learning Approach for Deep Face Recognition

Yandong Wen, Kaipeng Zhang, Zhifeng Li*, Yu Qiao
Shenzhen Institutes of Advanced Technology, CAS, China
The Chinese University of Hong Kong, Hong Kong, China

Introduction

- For generic object, scene or action recognition, the deeply learned features need to be separable. Because the classes of the possible testing samples are within the training set, the predicted labels dominate the performance.
- For face recognition task, the deeply learned features need to be not only separable but also discriminative. Since it is impractical to pre-collect all the possible testing identities for training, the label prediction in CNNs is not always applicable.
- The deeply learned features are required to be generalized enough for identifying new unseen classes without label prediction.

Discriminative Feature Learning

- SOFTMAX LOSS: encouraging the separability of features.
- CENTER LOSS: simultaneously learning a center for deep features of each class and penalizing the distances between the deep features and their corresponding class centers.
- JOINT SUPERVISION: minimizing the intra-class variations while keeping the features of different classes separable.

$$\mathcal{L} = \mathcal{L}_S + \lambda \mathcal{L}_C$$

$$\mathcal{L}_S = \sum_{i=1}^{m} \log \left( \sum_{j=1}^{m} e^{W^T \cdot x_i + b_j} \right) + \frac{\lambda}{2} \sum_{i=1}^{m} \| x_i - c_{y_i} \|^2$$

Experimental Results

- Labeled Face in the Wild (LFW) & Youtube Face (YTF)
  - The proposed model is trained on 7.0M face images, termed as model C.

<table>
<thead>
<tr>
<th>Method</th>
<th>Images</th>
<th>Networks</th>
<th>Acc. on LFW</th>
<th>Acc. on YTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFace [33]</td>
<td>4M</td>
<td>3</td>
<td>97.35%</td>
<td>94.1%</td>
</tr>
<tr>
<td>DeepID-2+ [32]</td>
<td>-</td>
<td>1</td>
<td>98.70%</td>
<td>-</td>
</tr>
<tr>
<td>FaceNet [27]</td>
<td>200M</td>
<td>1</td>
<td>99.47%</td>
<td>93.2%</td>
</tr>
<tr>
<td>Deep FR [25]</td>
<td>2.6M</td>
<td>1</td>
<td>98.85%</td>
<td>97.3%</td>
</tr>
<tr>
<td>Elite [21]</td>
<td>1.3M</td>
<td>1</td>
<td>99.13%</td>
<td>-</td>
</tr>
<tr>
<td>Model A</td>
<td>0.7M</td>
<td>1</td>
<td>97.37%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Model B</td>
<td>0.7M</td>
<td>1</td>
<td>99.10%</td>
<td>93.8%</td>
</tr>
<tr>
<td>Model C (Proposed)</td>
<td>0.7M</td>
<td>1</td>
<td>99.28%</td>
<td>94.9%</td>
</tr>
</tbody>
</table>

- MegaFace
  - Our model is trained on 490K face images, termed as model C*.

<table>
<thead>
<tr>
<th>Method</th>
<th>Protocol</th>
<th>Identification Acc. (Set 1)</th>
<th>Verification Acc. (Set 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NTechLAB - facenum_large</td>
<td>large</td>
<td>73.300%</td>
<td>85.081%</td>
</tr>
<tr>
<td>Google - FaceNet v8</td>
<td>Google - FaceNet v8</td>
<td>70.496%</td>
<td>86.473%</td>
</tr>
<tr>
<td>Beijing FaceAll Co. - FaceAllNorm_1600</td>
<td>Beijing FaceAll Co. - FaceAllNorm_1600</td>
<td>59.363%</td>
<td>59.036%</td>
</tr>
<tr>
<td>Barebones_FR - cnn</td>
<td>Barebones_FR - cnn</td>
<td>41.863%</td>
<td>41.297%</td>
</tr>
</tbody>
</table>

A Visualization Example on MNIST
Network of Experts for Large-Scale Image Categorization

Karim Ahmed, Mohammad Haris Baig, Lorenzo Torresani

Department of Computer Science, Dartmouth College, USA


Intuition

The visual system of a layperson is a very good generalist that can accurately discriminate coarse categories but lacks the specialist eye to differentiate categories that look alike.

Contribution

Inspired by this analogy, we propose a novel tree-structured architecture (Network of Experts) for large-scale image categorization. It can be built from any existing convolutional neural network (CNN), and the training is completely end-to-end.

Related Work

Our approach relates closely to methods that learn hierarchies of categories to train CNN experts, such as Hinton et al. [2], Warde-Farley et al. [7], and (HD-CNN) Yan et al. [8].

But it provides the following novel benefits:

- Improved accuracy, tested for 5 different base architectures.
- Does not require training of base model.
- Does not suffer from mistakes due to routing to the wrong expert.
- End-to-end optimization over the original categorization problem.

Technical Approach

We decompose large-scale image categorization into two separate tasks:

\[ \text{Given base model objective:} \]
\[ E_b(\theta; D) = R(\theta) + \frac{1}{N} \sum_{i=1}^{N} \left( \text{Image Class Label} - \text{Classification Loss} \right) \]

We propose the following generalist objective:

\[ E_g(\theta^g; D) = R(\theta^g) + \frac{1}{N} \sum_{i=1}^{N} \left( \text{Image Class Label} - \text{Classification Loss} \right) \]

where \( \theta^g \) maps classes to specialties

- \( i.e. \; \forall y \in \{1, \ldots, C\} \) assign \( (y_i) \in \{1, \ldots, K\} \) where \( \sum_{\# \text{Specialties}} = \sum_{\# \text{Classes}} \)

Minimized via alternation between:

1. Optimizing parameters \( \theta^g \) while keeping specialty labels fixed (traditional SGD).
2. Updating specialty labels \( \theta \) given the current estimate of weights \( \theta^g \).

Training the Network of Experts:

The Network of Experts is a tree-structured architecture:

- The trunk splits into \( K \) branches corresponding to the \( K \) learned specialties.
- Each branch is an expert optimized to distinguish the classes within its specialty.
- Final softmax layer over all \( C \) classes calibrates the outputs of the \( K \) experts.

Results

CIFAR100

<table>
<thead>
<tr>
<th>Model</th>
<th>K=2</th>
<th>K=5</th>
<th>K=10</th>
<th>K=20</th>
<th>K=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOF E</td>
<td>53.3</td>
<td>55.0</td>
<td>56.2</td>
<td>55.7</td>
<td>55.3</td>
</tr>
<tr>
<td>Base: AlexNet-C100</td>
<td>54.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Top-1 accuracy (%) on CIFAR100 for the base model (AlexNet-C100) and Network of Experts (NOF E) using varying number of experts (K).

ImageNet

<table>
<thead>
<tr>
<th>Approach</th>
<th>Top-1 # params</th>
<th>Avg. Inference time</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOF E using NIN</td>
<td>67.96</td>
<td>4.7M</td>
</tr>
<tr>
<td>HD-CNN [8] using NIN</td>
<td>67.38</td>
<td>9.2M</td>
</tr>
</tbody>
</table>

Table 2: Example of specialties learned from CIFAR100.

References

**Zero-Shot Recognition via Structured Prediction**

Ziming Zhang & Venkatesh Saligrama  
ECE, Boston University

### Experiments

**Exp I: Zero-Shot Recognition**

<table>
<thead>
<tr>
<th>Method</th>
<th>AP@Y</th>
<th>A@A</th>
<th>CUB</th>
<th>SUN</th>
<th>AVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Akata et al. [29]</td>
<td>34.10</td>
<td>52.80</td>
<td>57.20</td>
<td>57.09</td>
<td>46.05</td>
</tr>
<tr>
<td>Latour el. [14]</td>
<td>36.00</td>
<td>58.50</td>
<td>69.00</td>
<td>68.00</td>
<td>58.00</td>
</tr>
<tr>
<td>Zhang &amp; Saligrama [34]</td>
<td>40.00</td>
<td>59.00</td>
<td>59.00</td>
<td>59.00</td>
<td>60.00</td>
</tr>
<tr>
<td>BL-ZSL + SP-ZSR (This paper)</td>
<td>44.00</td>
<td>65.00</td>
<td>68.00</td>
<td>68.00</td>
<td>70.00</td>
</tr>
</tbody>
</table>

**Exp II: Zero-Shot Retrieval**

<table>
<thead>
<tr>
<th>Method</th>
<th>AP@Y</th>
<th>A@A</th>
<th>CUB</th>
<th>SUN</th>
<th>AVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL-ZSL + SP-ZSR (This paper)</td>
<td>32.00</td>
<td>52.00</td>
<td>52.00</td>
<td>52.00</td>
<td>52.00</td>
</tr>
</tbody>
</table>

**Previous Work: Key Insight**

- **Training:** Source domain is mapped to cluster center in target domain (or in embedded latent domain) (e.g. Akata et al. [29], Zhang & Saligrama [13, 34])

### This Paper: Transductive Zero-Shot Recognition

**Motivation**

- **a:** Observation: CNN features are quite reliable for supervised recognition, as they are clustered quite well for different classes.
- **b:** Assumption: There exist strong correlations (e.g. in terms of distance or similarity) between source domain data and cluster centers of target data distributions with different (seen and unseen) classes in latent spaces.

**Our Approach**

1. For each unseen test class, estimate its cluster center among unlabeled target data.
2. Then based on the cluster centers, update the assignments of target data instances as the predicted class labels with the help of unseen source domain data.
3. Repeat 1 and 2 until convergence.

#### Example

1. estimate cluster centers in target data
2. update assignments (i.e. predicted class labels) between target data and cluster centers

#### Evaluation

- Class-level precision comparison
- Class-level recall comparison
- Class-level AP comparison

**Reference:***


**Code Available:** [https://zimingzhang.wordpress.com/](https://zimingzhang.wordpress.com/)
A Generalized Successive Shortest Path Solver for Tracking Dividing Targets

Carsten Haubold, Janez Aleš, Steffen Wolf, Fred A. Hamprecht
HCI/IWR, University of Heidelberg

Introduction

- tracking-by-assignment is an important model for multi-target tracking
- often solved optimally by min-cost flow algorithms [Zhang, Pirsiavash, Lenz]
- when allowing for dividing targets (as in cell tracking), problem is NP-hard
- can be addressed with integer linear programming (ILP), but that does not scale
- we propose a network-flow based approximation with compelling anytime performance

Network Flow Setting

- we build a global model to incorporate temporal context akin to [Schiegg, Magnusson]
- assume imperfect segmentation: targets can merge and split, and spurious detections can happen
- set up a directed graph \( G=(V,E) \) as [Zhang], every unit of flow tracks one target
- detections represented as two nodes, connecting arc holds the detection cost
- arc costs \( v_{w(u,v), f(u,v)} \) are the current flow along \( (u,v) \)
- \( w(u,v,f(u,v)) \) must be convex w.r.t. \( f(u,v) \)
- arc capacities \( c(e) \) bounded by the maximally allowed mergers

Generalized Successive Shortest Paths (SSP)

- residual graph arc capacities defined as \( c_{r}(u,v) = c(u,v) - f(u,v), c(u,v) = f(u,v) \)
- Observation: residual arc capacity depends on the flow along the arc itself
- handle constraints by adjusting residual arc capacities between SSP iterations!
- allow residual arc capacities to depend on the flow along other arcs, e.g. \( c(u,v) = f(u,v) \)

Conditioned Residual Capacity Successive Shortest Paths (crcSSP) Algorithm:

- augment SSP by additional capacity update rule
- falls back to normal SSP when no constraints (e.g. divisions) are present
- add paths as long as they have negative cost, otherwise we would leave a minimum
- depending on the order of found paths and the capacity changes incurred by constraints, the optimal solution might be blocked
- greedy!

Residual Graph Approximations:

- [Magnusson]’s “swap arc” heuristic can now be understood as a second order residual graph approximation, where paths can only go backward for one frame at a time
- a first order approximation would mean no arcs pointing backward in time
- both can be solved by dynamic programming

Implementation Details:

- graph contains negative cost cycles \( \rightarrow \) use Bellman Ford (BF) to find shortest paths
- improve BF runtime:
  - add early stopping criteria to notice when a cycle is present:
    - check for cycle when source distance decreases or path longer than average
    - order edges by time before processing them with BF \( \rightarrow \)
  - reuse all computed distances that are still valid from the previous iteration \( \rightarrow \)

Experiments & Results

Datasets: (both with given ground truth)

a) Drosophila (from [Schiegg]), 100 time frames, 45K nodes, 10K division candidates
b) Pancreatic Rat Stem Cells (PSC) (from [Rapaport]), 104 time frames, 260K nodes and 770K arcs, 126K division candidates

Exemplary 2D slices of the 3D+t and 2D+t datasets, each showing from left to right:
1) raw data
2) raw data with resulting tracks, 10 frames before and after the current frame
3) our segmentation as additional overlay, where color indicates shared ancestry.

Anytime Performance:

- energy of solution after every new path using different SSP settings, compared to the optimal ILP energy
- 1st and 2nd order approximations are faster but lead to a larger optimality gap
- ordering arcs and only recomputing invalidated distances in BF is crucial
- warm-starting crcSSP from 1st or 2nd order approximations performs poorly
- crcSSP outperforms the ILP solver in runtime while using at most half the RAM
- is scalable: runtime complexity \( O(P^3 N^2) \), where \( P=\#tracks, N=\#nodes in graph \)

Tracking Quality:

- comparing the tracking results with the ground truth shows that, despite the heuristic nature our flow-based solver, the results are very close to the optimal ILP solution

Conclusion & Future Work

- we presented a new approximate primal feasible solver to handle side-constraints in network flow problems, such as division constraints in cell tracking
- in practice it gives close-to-optimal results, while providing iteratively improving intermediate solutions (useful e.g. for displaying interactive feedback)
- we want to apply this to other side-constraints in tracking

Source Code available at GitHub
- generalized SSP: https://github.com/chaubold/dpct
- ILP solver: https://github.com/chaubold/multiHypothesesTracking

References

Accurate and Linear Time Pose Estimation from Points and Lines

Alexander Vakhitov\(^1\), Jan Funke\(^2\) & Francesc Moreno-Noguer\(^2\)

\(^1\) St. Petersburg State University, St. Petersburg, Russia; Skolkovo Institute of Science and Technology, Moscow, Russia; alexander.vakhitov@gmail.com

\(^2\) Institut de Robòtica i Informàtica Industrial, UPC-CSIC, Barcelona, Spain, {jfunke,fmoreno}@irb.upc.edu

Camera Pose From Lines and Points

Point projection

\[
\pi_{\theta,X}(\theta, X) = \pi_{\theta}(X) = \sum_{i=1}^3 \alpha_i C_i
\]

Points cost:

\[
\|R \mu\|^2 \rightarrow \min
\]

Denote

\[
m_i(X) = \sum_{i=1}^3 \alpha_i C_i
\]

Point-to-line error:

\[
E_{\text{line}}(\theta, X) = \mu_i^T\mu_i
\]

Lines cost:

\[
E_{\text{line}} = \|\mu_i\|^2 + \|\mu_i\|^2
\]

Total cost:

\[
\|\mu_i\|^2 + E_{\text{line}} = \|R \mu\|^2 \rightarrow \min
\]

Contributions

- algebraic line error as linear constraints on the line endpoints
- \(\text{OPnP} \rightarrow \text{EPnP} \rightarrow \text{EPnP} \)
- Key features:
  - camera pose accuracy increased
  - negligible runtime overhead

Synthetic Experiments

(a) Accuracy w.r.t. noise level, parallel lines: \(\sigma = 0, \text{noise} = 10\)

(b) Accuracy w.r.t. feature type, parallel lines: \(\sigma = 1\)

(c) Accuracy w.r.t. noise level, nonparallel lines

(d) Running time of the algorithms on the planes of OPnP, EPnP

Real Experiments

(a) OPnP (pt)

(b) OPnP (lin)/(pt+lin)

(c) OPnP (pt)

(d) OPnP (lin)/(pt+lin)

(Ref. Err. (pts), Translation Error (m))


Objective: A method that is able to rapidly and robustly determine the projective spacing model from a minimum set of assigned line combinations.

Prior Art: Let a point in homogeneous form be \( p=(x,y,z) \) and line \( l=(a,b,c) \). A group of equally spaced parallel lines can be represented as \( ax+by+\lambda = 0 \), where \( \lambda = 0, 1, \ldots, n \) (is a scalar index of a line) takes only integer values. Under perspective imaging the line transformation is [1]

\[
I_k = H^{-T} \begin{bmatrix} 0 \\ \lambda \end{bmatrix} = H^{-T} \begin{bmatrix} a \\ b \\ 0 \end{bmatrix} \begin{bmatrix} 1 \\ \lambda \end{bmatrix} = A \begin{bmatrix} 1 \\ \lambda \end{bmatrix}
\]

(1)

where \( H \) is the homography matrix and \( A = [I_k | I_{\infty}] \) is a \( 3 \times 2 \) matrix. This leads to vector cross product [2] as follows

\[
I_k \times A \begin{bmatrix} 1 \\ \lambda \end{bmatrix} = 0
\]

(2)

which can be arranged in \( Z_kx_k = 0 \) format and solved using Singular Value Decomposition (SVD) for \( x_k \), i.e., \( I_k \) and \( I_{\infty} \) [2].

Pseudo-Geometric Formulation:

We know from Eqn. 1 that

\[
I_k = I_0 + \lambda I_{\infty}
\]

(3)

First, we rework this formulation to interpolate between real lines avoiding the use of line at infinity \( I_{\infty} \). For line \( n \) we get

\[
I_n = I_0 + nI_{\infty}
\]

(4)

Re-arranging gives us

\[
I_0 = (I_n - I_0)/n
\]

(5)

Substituting Eqn. 5 in Eqn. 3 to eliminate \( I_{\infty} \) to get

\[
I_k = (n-\lambda)n_0x_k + \lambda I_{\infty}
\]

(6)

Second, we exploit the dot product between the interpolated lines of the equidistant projective spacing model and points on edge lines in the image, i.e.,

\[
I_k \cdot p = 0
\]

\[
(n-\lambda)n_0x_k + \lambda I_{\infty} \cdot (x,y,1) = 0
\]

\[
(n-\lambda)n_0x_k + \lambda b_0y + (n-\lambda)c_0 + \lambda a_0 + \lambda b_0 + \lambda c_0 = 0
\]

and can be arranged in \( 2x = 0 \) format as

\[
\begin{bmatrix}
(n-\lambda)x_{k,1} & (n-\lambda)y_{k,1} & x_{k,1} & y_{k,1} & \lambda_1 & \cdots \\
(n-\lambda)x_{k,2} & (n-\lambda)y_{k,2} & x_{k,2} & y_{k,2} & \lambda_1 & \cdots \\
(n-\lambda)x_{k,3} & (n-\lambda)y_{k,3} & x_{k,3} & y_{k,3} & \lambda_1 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\
(n-\lambda)x_{k,M} & (n-\lambda)y_{k,M} & x_{k,M} & y_{k,M} & \lambda_1 & \cdots \\
\end{bmatrix} = 0
\]

(7)

and solved using SVD for \( x \), i.e., \( I_k = (a_0, b_0, c_0) \) and \( I_{\infty} = (a_0, b_0, c_0) \).

Experimental Results:

Figure 1. Visual comparison of equidistant model parallel (colour dashed) lines overlayed on all the image (bold colour and black) lines: (a) our work, (b) baseline, (c) our work optimized, and (d) baseline optimized, when fitted using three image lines (black lines).

Conclusion:

We proposed a linear pseudo-geometric formulation which is based on the dot product of interpolated model lines against end points in the image for equidistant parallel line fitting under perspective distortion. Simulated and real data experiments showed that our improved solution does not require any pre-conditioning of the image data and avoids the need for computationally expensive non-linear optimization.

References:

Towards perspective-free object counting with deep learning

Daniel Oñoro-Rubio and Roberto J. López-Sastre
GRAM, University of Alcalá

Understanding the counting problem

- Given an image and a target object, we want to guess the total number of objects.

Our Counting CNN is a fully convolutional model that maps an input image patch into its estimated density map.

The counting by regression model with deep learning

\[ D_i(p) = \sum_{\mu \in A_i} N(p; \mu, \Sigma) \]

\[ n^o \]

\[ N_i = \sum_{p \in I} D_i(p) \]

Our models

- **Counting CNN** uses a pyramid of input patches cropped from the center of the target patch to provide multiscale information to the network.

Experiments

**TRANCOS**
- Target object: Cars
- Different scenes
- No perspective map
- 1244 images

**UCSD**
- Target object: People
- Single scene
- With perspective map
- 2000 images
- Video sequence

**UCF**
- Target object: People
- Multiple scenes
- No perspective map
- 50 images
- Count between 94-4543
- Average of 1280 per image

Our models

- **Counting CNN**
  - **BASE**
  - **CCNN**
  - **HYDRA 2S**
  - **HYDRA 3S**
  - **HYDRA 4S**
  - **HYDRA 4S**

Table of Results

<table>
<thead>
<tr>
<th>METHOD</th>
<th>GAME 0</th>
<th>GAME 1</th>
<th>GAME 2</th>
<th>GAME 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIASCHI ET AL.</td>
<td>17.77</td>
<td>20.14</td>
<td>23.65</td>
<td>25.99</td>
</tr>
<tr>
<td>LEMPITSKY ET AL.</td>
<td>13.76</td>
<td>16.72</td>
<td>20.72</td>
<td>24.36</td>
</tr>
<tr>
<td>CCNN</td>
<td>12.49</td>
<td>16.58</td>
<td>20.02</td>
<td>22.42</td>
</tr>
<tr>
<td>HYDRA 25</td>
<td>11.41</td>
<td>16.36</td>
<td>20.89</td>
<td>23.67</td>
</tr>
<tr>
<td>HYDRA 35</td>
<td>10.99</td>
<td>13.75</td>
<td>16.69</td>
<td>19.32</td>
</tr>
<tr>
<td>HYDRA 45</td>
<td>12.92</td>
<td>15.54</td>
<td>18.45</td>
<td>20.96</td>
</tr>
</tbody>
</table>

Table of Results

<table>
<thead>
<tr>
<th>METHOD</th>
<th>MAXIMAL</th>
<th>DOWNSCALE</th>
<th>UPSCALE</th>
<th>MINIMAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEMPITSKY ET AL.</td>
<td>1.70</td>
<td>1.28</td>
<td>1.99</td>
<td>2.02</td>
</tr>
<tr>
<td>FIASCHI ET AL.</td>
<td>1.70</td>
<td>2.16</td>
<td>1.61</td>
<td>2.20</td>
</tr>
<tr>
<td>PHAM ET AL.</td>
<td>1.43</td>
<td>1.30</td>
<td>1.59</td>
<td>1.62</td>
</tr>
<tr>
<td>ARTETA ET AL.</td>
<td>1.24</td>
<td>1.31</td>
<td>1.69</td>
<td>1.49</td>
</tr>
<tr>
<td>ZHANG ET AL.</td>
<td>1.70</td>
<td>1.26</td>
<td>1.59</td>
<td>1.52</td>
</tr>
<tr>
<td>CCNN</td>
<td>1.65</td>
<td>1.79</td>
<td>1.11</td>
<td>1.50</td>
</tr>
<tr>
<td>HYDRA 25</td>
<td>2.22</td>
<td>1.93</td>
<td>1.37</td>
<td>2.38</td>
</tr>
<tr>
<td>HYDRA 35</td>
<td>2.17</td>
<td>2.99</td>
<td>1.44</td>
<td>1.92</td>
</tr>
</tbody>
</table>

Table of Results

<table>
<thead>
<tr>
<th>METHOD</th>
<th>MAE</th>
<th>MSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RODRIGUEZ ET AL.</td>
<td>465.7</td>
<td>697.8</td>
</tr>
<tr>
<td>LEMPITSKY ET AL.</td>
<td>493.4</td>
<td>487.1</td>
</tr>
<tr>
<td>ZHANG ET AL.</td>
<td>467.0</td>
<td>498.5</td>
</tr>
<tr>
<td>IDRIES ET AL.</td>
<td>419.5</td>
<td>541.6</td>
</tr>
<tr>
<td>ZHANG ET AL.</td>
<td>377.6</td>
<td>529.1</td>
</tr>
<tr>
<td>CCNN</td>
<td>488.67</td>
<td>646.68</td>
</tr>
<tr>
<td>HYDRA 25</td>
<td>333.73</td>
<td>425.26</td>
</tr>
<tr>
<td>HYDRA 35</td>
<td>465.73</td>
<td>371.84</td>
</tr>
</tbody>
</table>

This work is supported by the projects SPIP2014-1468, SPIP2015-01809, and MINECO TEC2013-45183-R.
Information Bottleneck Domain Adaptation with Privileged Information for Visual Recognition

Saeid Motiian and Gianfranco Doretto
Lane Department of Computer Science and Electrical Engineering, West Virginia University

Introduction

The information bottleneck (IB) method \[2\] is a technique used to find a compact representation of the input data that retains as much information as possible about a target variable. However, in some cases, we may have additional information that is not available at test time, which is called privileged information. In such cases, we can use the IB method with privileged information (IBDAPI) to adapt to the test domain.

Domain Adaptation with Privileged Information

IB domain adaptation with privileged information (IBDAPI)

\[ C_{d}(S|X) = \beta C_{d}(S|X, S) - \gamma C_{d}(Y|X, S) \]  \( (1) \)

\[ C_{d}(S|Y) = \beta C_{d}(S|Y, S) - \gamma C_{d}(Y|Y, S) \]  \( (2) \)

Learning A and W

IBDAPI for visual recognition

\[ \min_{A, W} C_{d}(X, A) + \lambda C_{d}(X, W) \]  \( (3) \)

Large-margin IBPI

In order to develop a large-margin classifier, we assume:

- \( f = q(k, \omega) \) and \( T = q(k, \omega) \), where \( k \) is a suitable set of parameters.
- \( f(k) \) is a large margin decision function given by \( f = \arg \min_{w} \frac{1}{2} \|w\|^2 + \frac{1}{2} \sum_{i=1}^{\infty} \lambda \|w_{i}\|^2 \) \( (4) \)

Experiments

Object Recognition: RGB-D Object Dataset \( \Rightarrow \) Caltech256 Dataset

USA-Office Dataset

Conclusion Recognition: EURECOM Dataset \( \Rightarrow \) LFW-a Dataset

References

Overview

We introduce a template-free approach to reconstructing a poorly-textured, deformable surface. To this end, we leverage surface isometry and formulate 3D reconstruction as the joint problem of non-rigid image registration and depth estimation. Our experiments demonstrate that our approach yields much more accurate 3D reconstructions than state-of-the-art techniques.

Motivation

- Existing NRSfM approaches have only been applied to relatively well-textured objects and are ill-suited to handle the challenging case of poorly-textured surfaces.
- Template-based nonrigid 3D reconstruction methods all require a known 3D template of the surface, which, in many practical situations, is difficult, if not impossible, to obtain.
- We aim to achieve the best of both worlds: We introduce an approach that does not require a template of the surface of interest, but can nonetheless reconstruct poorly-textured surfaces.

Our Approach

We formulate 3D reconstruction as the problem of jointly estimating the 2D displacement of each mesh vertex in each frame, except for the first frame, and the depth of the vertices in all the frames. This is expressed as an optimization problem containing an image-based energy and shape priors.

Image-based Energy

Since we aim at tackling the poorly-textured scenario, we cannot expect to be able to reliably track feature points across the frames. Instead, we therefore exploit two sources of image information.

\[ E_I(U) = E_I(U) + w_E E_E(U) \]

Brightness constancy. Our first image-based term relies on the intuition that the intensity under a mesh point should remain constant as the mesh deforms.

\[ E_I(U) = \frac{1}{F-1} \sum_{f=2}^{F} \sum_{n=1}^{N} ||I_f^i(x_n^f) - I_f^i(x_n^0)||^2 \]

Image Edges. We further account for the image edges, which, while sparse in the poorly-textured case provide more reliable information.

\[ E_E(U) = \frac{1}{F-1} \sum_{f=2}^{F} \sum_{n=1}^{N} ||D(x_n^f)||_2 \]

Shape Priors

Relying on image information only is known to leave many ambiguities in non-rigid shape reconstruction. Here, we make use of three such priors: Isometry, spatial smoothness and temporal smoothness.

\[ E_{\text{shape}}(U, D, L) = w_I E_I(U, D, L) + w_E E_E(U, D) + w_{\text{const}} E_{\text{const}}(U, D) \]

Isometry. This prior enforces the fact that the distance between two neighboring 3D points should not vary, or vary minimally, as the surface deforms. Here, we encourage this for every edge in our mesh.

\[ E_{\text{shape}}(U, D, L) = \frac{1}{F} \sum_{f=2}^{F} \sum_{n=1}^{N_c} \sum_{j \in \mathcal{N}_n} (||v_n^f - v_j^f||_2 - l_{n,j})^2 \]

Spatial Smoothness. We also rely on the intuition that the shape of the surface remains relatively smooth as it deforms.

\[ E_{\text{shape}}(U, D) = \frac{1}{F} \sum_{f=2}^{F} \left( \sum_{n=1}^{N} \sum_{j \in \mathcal{N}_n} (||D_n^f||_2^2 + \frac{1}{|T|} \sum_{n \in T} (||D_n^f||_2^2 + ||D_n^f||_2^2) \right) \]

Temporal Smoothness. As the input is a video sequence, we model the natural intuition that sudden changes in our parameters are unlikely to occur between neighboring frames.

\[ E_{\text{shape}}(U, D) = \frac{1}{F-1} \sum_{f=2}^{F} \sum_{n=1}^{N} ||D_n^f||_2^2 \]

Fusion-moves

Since this is a non-convex problem, we make use of a fusion moves strategy to optimize it, which has proven more effective than gradient-based optimization in practice.

Proposal Generation

Given the current solution, we firstly update the image displacements by 2D tracking, and then update the depths by minimizing the following equation.

\[ D^* = \arg \min_D E_{\text{shape}}(U^*, D^*, L^*) \]

Experimental Results

Baselines

Vicente12: This method corresponds to the template-free approach, which leverages isometry in an NRSfM context.

Garg13: This method relies on a total variation regularization within a dense NRSfM framework based on the optical flow.

Chakthul14: This method performs template-free 3D shape reconstruction by relying on isometry and on an infinitesimal planarity assumption.

Quantitative Evaluation

Cardboard College Cardboard Minimal texture College Cloth

Experimental Results

Real Images Ours Vicente12 Garg13 Chakthul14

These results clearly evidence that our approach yields much more accurate results than the baselines. In particular, the baselines all suffer from the lack of texture and the illumination changes, which make the feature-matching, or 2D registration, step unreliable. By contrast, our approach that jointly performs 2D registration and 3D reconstruction essentially regularizes the matching problem with 3D constraints, and thus yields much more accurate results.

Conclusion

We have formulated reconstruction as the problem of jointly optimizing the 2D image displacements of mesh vertices and the depth of these vertices, and have proposed a fusion moves strategy to optimize the resulting problem. Our experiments have demonstrated the effectiveness of our approach, and have shown that it yields much higher accuracy than existing template-free techniques. To the best of knowledge, this constitutes the first attempt at solving the challenging template-free and poorly-textured scenario.

Future Works. We intend to study solutions to address the case where the first frame depicts large, complex deformations, which our current approach remains ill-suited to handle.

Acknowledgement: This work was supported in part by Natural Science Foundation of China (No.61231018, No.61273366), National Science and Technology Support Program (2015BAH31F01) and Program of Introducing Talents of Discipline to University under grant B13043. Part of this work was performed while X. Wang and M. Salzmann were respectively visiting and affiliated with NICTA, Canberra.
FigureSeer
Parsing Result-Figures in Research Papers

Noah Siegel  Zachary Horvitz  Roie Levin  Santosh Divvala  Ali Farhadi

Paper corpus from www.semanticscholar.org

Figure Extraction using www.pdffigures.allenai.org

Figure Classification using CNN trained on 60K figures

Input: All Extracted ‘Graph’ Figures

Axis Detection (Position, Label, Scale)

Legend Detection (Labels, Symbols)

Plot Data Detection (Optimal Path Tracing)

Output: Structured Data

Path Tracing Approach

Find a path \( P_a = \{(x_i)_i \}_{i=1}^n \) maximizing:
\[
E(P_a) = \sum_{i=1}^{n} \phi_i(x_i) + \sum_{i=1,j=i+1}^{n} \phi_{ij}(x_i, x_j),
\]
\( s.t. \), \( 1 \leq y \leq m, \ 1 \leq x_i \leq n, \ x_{i+1} = x_i + 1. \)

Learning \( \phi \):
Randomly initialize \( \phi \)
Until convergence:
Compute \( P_a \) under \( \phi \)
Set \( \phi \) to maximize \( E(P^*)-E(P_a) \) (rank SVM)

Path Tracing Features

Qualitative Analysis

Key Challenges: Strict Requirements; High Variability; Heavy Clutter

Quantitative Results

With Task-specific Evaluation Metrics

<table>
<thead>
<tr>
<th>Task</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure Extraction</td>
<td>90%</td>
</tr>
<tr>
<td>Figure Classification</td>
<td>86%</td>
</tr>
<tr>
<td>Axes Detection</td>
<td>91%</td>
</tr>
<tr>
<td>Legend Detection</td>
<td>72%</td>
</tr>
<tr>
<td>Plot Data Detection</td>
<td>26%</td>
</tr>
<tr>
<td>Query Answering</td>
<td>47%</td>
</tr>
</tbody>
</table>

Key Application: Query Answering

Powerful search & querying of complex figure content across multiple papers

www.allenai.org/plato/figureseer
**Goal**

Approximate nearest neighbor (ANN) search

**I. Contributions**
- A novel quantization approach based on constrained sparse coding for ANN search.
- The proposed approach is demonstrated with an adaptation of two vector quantizers:
  - Product Quantizer (PQ) as "Qα-PQ"
  - Residual Vector Quantizer (RVQ) as "Qα-RVQ"
- Competitive results on four datasets including billion-sized BIGANN dataset.
- Significant improvement over RVQ.
- Faster encoding with similar or better accuracy.

**II. Background**
- ANN search attempts to find a query’s nearest neighbor efficiently by relying on compressed representations.
- A vector quantizer (VQ) maps a vector to a codeword, $x \mapsto c \in C$.
  $$Q(x) = \operatorname{arg~min}_{c \in C} \|x - c\|_2$$

- Product quantizer (PQ) [2]
  - An encoded vector takes $M \times \log K$ bits.
- Residual vector quantizer (RVQ) [1]
  - The sub-codebooks form a hierarchy and each represents a different level of residuals.

**III. Approach**
- **Proposed representation**
  $$Q(x) = \sum_{m=1}^{M} a_m(x) c_m^T(x)$$
  $$a_m(x) = \alpha \cdot c_m^T(x)$$

- Coefficient vectors formed as $\alpha = [\alpha_1(x), \ldots, \alpha_M(x)]$ are vector quantized.

- **PQ**
  - $Q(x) = \max_{c \in C} \langle x, c \rangle$
  - $x$ is encoded as $(k_1, \ldots, k_M)\alpha$ where $k_m = \operatorname{arg~max}_{c \in C} \langle c_m, x \rangle$.
- **RVQ**
  - $Q(x) = \max_{r \in R} \langle r, x \rangle$
  - $x$ is encoded as $(k_1, \ldots, k_M)\alpha$ where $k_m = \operatorname{arg~max}_{r \in R} \langle r, c_m \rangle$.

**REFERENCES**

4. All about VLAD. Arandjelovic, Raja and Zisserman, Andrew. CVPR. 2013.
8. Optimized residual vector quantization for efficient approximate nearest neighbor search. As, Liefu and Yu, Junqing and Wu, Zhibin and He, Yantong and Guan, Tao Multimedia System. 2015.

**IV. Results: ANN Search**
- **Impact of 1-byte quantization of $\alpha$**
  - Accuracy with quantized $\alpha$ compared to unquantized $\alpha$ and the associated VQ.
  - $Qα-RVQ$ retains the accuracy of $α-RVQ$.
  - $Qα-PQ$ performs well for lower $M$.

- **Comparison with state-of-the-art**
  - Accuracy of various quantization methods using the same compression rate.
  - $Qα-RVQ$ performs the best after additive quantization(AQ). [6], a highly accurate but not scalable approach.

- **Efficiency**
  - Comparing complexity of codebook learning, database encoding and ANN search.

  In the above experiments, we set $P = 256$ and $K = 256$, except for $Qα-PQ$ in Fig. 2 and Tab. 1 where $K=128$.

<table>
<thead>
<tr>
<th>Method</th>
<th>PQ</th>
<th>RVQ</th>
<th>Qα-RVQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>learn 1</td>
<td>0.212</td>
<td>1.250</td>
<td>0.719</td>
</tr>
<tr>
<td>encode 1</td>
<td>0.206</td>
<td>1.342</td>
<td>0.615</td>
</tr>
<tr>
<td>search 1</td>
<td>1.867</td>
<td>1.225</td>
<td>1.950</td>
</tr>
</tbody>
</table>

**V. Results: Non-exhaustive ANN search**

- **Results on Billion-sized BIGANN** [Jegou et al., ICASSP 2011] dataset using two inverted indexing methods.

- **Inverted index** [2] uses a coarse quantizer and then the residual is quantized using the chosen encoder.
  - $Qα-RVQ$ is the most accurate method.
  - The hierarchical structure of $Qα-RVQ$ permits to further narrow the exhaustive search leading to much faster search.

- **Inverted multi-index** [3] uses two-fold product quantizers with a large number of inverted lists.
  - $Qα-RVQ$ and $Qα-PQ$ outperform others in accuracy.
  - Interestingly $Qα-RVQ$ can be more accurate and faster than PQ by scanning fewer candidates.

**TABLE 1**: Complexity w.r.t. PQ on SIFT1M (64-bit encoding).

<table>
<thead>
<tr>
<th>Method (bytes)</th>
<th>PQ-RVQ</th>
<th>Qα-RVQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 100K$</td>
<td>0.111</td>
<td>0.288</td>
</tr>
<tr>
<td>$T = 300K$</td>
<td>0.124</td>
<td>0.471</td>
</tr>
</tbody>
</table>

**TABLE 2**: Performance on BIGANN with relative timings.

<table>
<thead>
<tr>
<th>Method (bytes)</th>
<th>PQ-RVQ</th>
<th>Qα-RVQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T = 100K$</td>
<td>0.139</td>
<td>0.450</td>
</tr>
<tr>
<td>$T = 300K$</td>
<td>0.160</td>
<td>0.514</td>
</tr>
</tbody>
</table>

**TABLE 3**: Performance with $T$ candidates exhaustively scanned. Time is in milliseconds.
Sympathy for the details: Dense trajectories and hybrid classification architectures for action recognition

César Roberto de Souza\textsuperscript{1,2}, Adrien Gaidon\textsuperscript{1}, Eleonora Vig\textsuperscript{3}, Antonio Manuel López\textsuperscript{2}

\textsuperscript{1}Xerox Research Centre Europe, France; \textsuperscript{2}Computer Vision Center & Univ. Autònoma de Barcelona, Spain; \textsuperscript{3}German Aerospace Center, Germany

{cesar.desouza, adrien.gaidon}@xrce.xerox.com, eleonora.vig@dlr.de, antonio@cvc.uab.es - https://arxiv.org/abs/1608.07138

Context
• Action recognition in videos: shallow or deep?

Learned features
• State-of-the-art in many tasks
• Requires large annotated data sets

Hybrid models: best of both worlds
• State-of-the-art action recognition performance
• Data efficient (trained on 150 to 10,000 samples)

Handcrafted features
• Shallow models / simpler training
• Well stablilshed and tested pipeline

Hybrid supervised/unsupervised, handcrafted/deep/classification architecture

Input video

Unsupervised representation layers

Feature stacking
Low-level descriptors

Unsupervised

Feature stacking
Low-level descriptors

Unsupervised

Data augmentation and feature extraction

Dimensionality reduction layers

Supervised layers

Hybrid architecture ranking

Creating a strong baseline (best practices that work across datasets)

Paying careful attention to details
• Data augmentation (a common component in deep learning)
• RootSIFT normalization followed by PCA-whitening
• Spatio-temporal augmentation (STP vs. SFV vs. STA)
• FV-level fusion with double Power normalization
• Supervised dimensionality reduction
• PCA up to 99% variance explained + whitening + $\ell_2$
• Batch normalization (BN), ReLU (RL), Dropout (DO)

Data Augmentation by Feature Stacking (DAFS)
• Extension of MIFS [7] for arbitrary transformations
• Horizontal flipping (video mirroring)
• Frame-skipping (level 0, 1, 2 and 3)
• Aggregates features (FVs) instead of predictions
• Applicable for shallow, deep, hybrid models
• More efficient than standard data augmentation:
  • No increase in number of training samples

Creating data efficient models (comparison with the state of the art)

Examples of success and failure cases of our hybrid architecture compared to shallow models

References

Human pose estimation via convolutional part heatmap regression

Adrian Bulat and Georgios Tzimiropoulos

Computer Vision Laboratory, School of Computer Science, University of Nottingham, UK

Motivation

- Human pose estimation remains a very challenging problem
- Current methods fail to effectively deal with occlusion.

Contributions

- A simple CNN cascade that first detects and then refines the parts’ location through joint regression of part detection heatmaps
- Efficient occlusion handling by the detection CNN that guides the network where to focus
- We show the benefit of the detection-followed-by-regression approach with both VGG-FCN [6] and residual [4] architectures
- Top performance on both LSP and MPII datasets

Analysis

- We compare detection only, regression only, 2-step regression with the proposed detection-followed-by-regression approach
- Both VGG-FCN and residual architectures were considered

Network architecture

Sub-networks specifications


Comparison with state-of-the-art on MPII

PCKh-based comparison on MPII, for a full list please check http://human-pose.mpi-inf.mpg.de/#results

References

1. Motivation

- Intermediate features learned at different layers in a CNN are suitable for discriminating objects of different complexities.
- In the training phase, if a sample has already been correctly classified at a specific layer with high confidence, the rest layers should focus on classifying other difficult samples.

2. Main idea

- First, introduce multiple classifiers on top of multiple layers.
- Second, each classifier coordinates with other classifiers to jointly maximize the final classification performance.

3. CLDL

(A) Cross-layer heterogeneities

![Image 1: Three classifiers introduced at bottom/top layers of a deep model can correctly classify simple/complex samples.]

(B) Architecture of CLDL-DNN

\[ \mathcal{L}(\mathbf{x}, \mathbf{y}^*, \mathcal{W}) = \sum_{m=1}^{M} \lambda_m \phi^{(m)} + \alpha \| \mathcal{W} \|_2 \]

Inference strategy

\[ y^* = \arg \max_y \mathcal{L}(\mathbf{x}, \mathbf{y}^*, \mathcal{W}) \]

(C) Explanation of CLDL

- \( T^{(m)} \) measures how well those “companion” classifiers perform on classifying the input sample
- Consider \( C^{(m)} \) together with \( T^{(m)} \) distinguishes CLDL loss from conventional loss: each classifier considers the performance of other classifier when trying to classify input
- CLDL can be viewed as a simplified version of conditional random field (CRF) model. Please refer to proof in paper.

4. Results

- The effect of classifier number in CLDL on the classification accuracy

![Image 2: Evaluation of NIN model on CIFAR-100 with different number of classifiers in CLDL. Left: without data augmentation. Right: with data augmentation]

- Result for object classification

  - For CIFAR-100 and MNIST, NIN are used as baseline
  - For ImageNet, GoogleLeNet is used as baseline
  - Three classifiers are used in CLDL

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Maxout</th>
<th>NIN</th>
<th>DSN</th>
<th>GoogleLeNet</th>
<th>CLDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR-100</td>
<td>38.57</td>
<td>35.96</td>
<td>34.57</td>
<td>-</td>
<td>30.40</td>
</tr>
<tr>
<td>CIFAR-100*</td>
<td>-</td>
<td>32.75</td>
<td>-</td>
<td>-</td>
<td>29.05</td>
</tr>
<tr>
<td>MNIST</td>
<td>0.47</td>
<td>0.42</td>
<td>0.39</td>
<td>-</td>
<td>0.28</td>
</tr>
<tr>
<td>ImageNet</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>11.1</td>
<td>10.21</td>
</tr>
</tbody>
</table>

*with data augmentation

- Result for scene classification

  - VGGNet with 11/16/11 convolutional layers are used as baseline for MIT67, SUN397, Places205, respectively.
  - Three classifiers are used in CLDL

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT67</td>
<td>54.32</td>
<td>70.80</td>
<td>77.50</td>
<td>68.20</td>
<td>76.10</td>
<td></td>
</tr>
<tr>
<td>SUN397</td>
<td>68.24</td>
<td>51.90</td>
<td>56.20</td>
<td>54.30</td>
<td>60.70</td>
<td></td>
</tr>
<tr>
<td>Places205</td>
<td>50.00</td>
<td>-</td>
<td>-</td>
<td>80.90</td>
<td>84.10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT67</td>
<td>54.32</td>
<td>70.80</td>
<td>77.50</td>
<td>68.20</td>
<td>76.10</td>
<td>84.69</td>
</tr>
<tr>
<td>SUN397</td>
<td>68.24</td>
<td>51.90</td>
<td>56.20</td>
<td>54.30</td>
<td>60.70</td>
<td>70.40</td>
</tr>
<tr>
<td>Places205</td>
<td>50.00</td>
<td>-</td>
<td>-</td>
<td>80.90</td>
<td>84.10</td>
<td>88.67</td>
</tr>
</tbody>
</table>

For CIFAR-100 and MNIST, NIN are used as baseline.
For ImageNet, GoogleLeNet is used as baseline.
Three classifiers are used in CLDL.
**Goal**
Segment dynamic objects given a single space-variantly blurred image of a 3D scene captured using a hand-held camera.

**Challenges**
- Single Image
- Camera/object motion \(\rightarrow\) motion blur
- 3D scene \(\rightarrow\) defocus blur
- General camera motion/3D scene \(\rightarrow\) space-varying blur
- Depth-motion ambiguity

**Our approach**
Train a CNN to predict the composite kernel \(h_0\) at each pixel.

- Composite kernel is convolution of defocus \(h_d\) and motion \(h_m\) kernels.
- Use defocus cue to recover the depth map.
- Use motion kernels to segregate the dynamic objects at each depth layer.
- Joint model for defocus and motion helps resolve depth-motion ambiguity.

**Scene segmentation**
Layer with maximum area in depth map = Reference depth layer \(d_0\)

**Segmenting moving objects in the reference depth layer \(d_0\)**
Blur on dynamic object pixel \# Blur on pixel affected only by camera motion

**Kernel classification using CNN**

**Results**

**References**
**Highlights**

- Cast video summarization as a supervised structured prediction problem
- Use LSTM to model the variable-range dependencies among video frames for generating representative and compact summaries
- Exploit auxiliary datasets to enhance the performance via domain adaptation

**Approach**

**Modeling variable-range dependencies**

- vsLSTM: applying LSTM to capture and adjust long-term structural dependencies in a data-driven way

**Modeling diversity**

- Determinantal Point Process (DPP): defines the probability of selecting a subset \( y \) from an N-item ground set. Given the similarity kernel \( L \), diverse and representative subsets are highly probable:

  \[
  P(Y = y) = \frac{\det(L(y))}{\det(L + I)} \quad P(Y = (i,j)) \propto \det \left( \frac{L_{ij}}{L_{ij} + L_{ii}} \right)
  \]

- dppLSTM: learn to model pairwise frame-level "repulsiveness"

**Experiments**

**Qualitative results**: dppLSTM (red) and MLP-shot (green)

Positive example: dppLSTM captures the sequential flow to identify the main event

Negative example: dppLSTM fails to select the temporally-crowded high score frames

**Comparisons w/ state-of-the-art methods**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>C</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumMe</td>
<td>Zhang et al. 2016</td>
<td>40.9</td>
<td>41.3</td>
<td>38.5</td>
</tr>
<tr>
<td></td>
<td>vsLSTM (ours)</td>
<td>37.6</td>
<td>41.6</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>dppLSTM (ours)</td>
<td>38.6</td>
<td>42.9</td>
<td>41.8</td>
</tr>
<tr>
<td></td>
<td>Song et al. 2015</td>
<td>50.0</td>
<td>57.9</td>
<td>56.9</td>
</tr>
<tr>
<td></td>
<td>vsLSTM (ours)</td>
<td>54.2</td>
<td>57.9</td>
<td>56.9</td>
</tr>
<tr>
<td></td>
<td>dppLSTM (ours)</td>
<td>54.7</td>
<td>59.6</td>
<td>58.7</td>
</tr>
</tbody>
</table>

**Comparisons w/ MLP baselines**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>C</th>
<th>A</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumMe</td>
<td>MLP-shot</td>
<td>39.8</td>
<td>40.7</td>
<td>39.8</td>
</tr>
<tr>
<td></td>
<td>MLP-Frame</td>
<td>38.2</td>
<td>41.2</td>
<td>40.2</td>
</tr>
<tr>
<td></td>
<td>vsLSTM</td>
<td>37.6</td>
<td>41.6</td>
<td>40.7</td>
</tr>
<tr>
<td></td>
<td>vsLSTM</td>
<td>55.2</td>
<td>56.7</td>
<td>55.5</td>
</tr>
<tr>
<td></td>
<td>vsLSTM</td>
<td>54.2</td>
<td>57.9</td>
<td>56.9</td>
</tr>
</tbody>
</table>

**Domain Adaptation reduces cross-dataset discrepancies**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Transfer w/o DA</th>
<th>Transfer w/ DA</th>
<th>Augmented w/o DA</th>
<th>Augmented w/ DA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SumMe</td>
<td>vsLSTM</td>
<td>40.7</td>
<td>41.3</td>
<td>41.6</td>
<td>42.1</td>
</tr>
<tr>
<td></td>
<td>dppLSTM</td>
<td>41.8</td>
<td>43.1</td>
<td>42.9</td>
<td>44.7</td>
</tr>
<tr>
<td></td>
<td>vsLSTM</td>
<td>56.9</td>
<td>57.0</td>
<td>57.9</td>
<td>58.0</td>
</tr>
<tr>
<td></td>
<td>dppLSTM</td>
<td>58.7</td>
<td>58.9</td>
<td>59.6</td>
<td>59.7</td>
</tr>
</tbody>
</table>

**Introduction**

Indispensable video summarization

> 500 hrs new Youtube video per min

Popular ways: key frame shot selection

**Motivation**

Whether a frame is representative or not depends on other frames on the temporal line of the video

Existing works

Most don’t model the temporal information of video frames or merely model within fixed-length intervals.

**Contribution**

Modeling variable-range dependencies where both short-range and long-range relationships intertwine.

**Video Summarization with Long Short-term Memory**

Ke Zhang*1, Wei-Lun Chao*1, Fei Sha2, and Kristen Grauman3

1University of Southern California, 2University of California, Los Angeles, 3University of Texas at Austin

K.G. is partially supported by NSF IIS-1514118 and a gift from Intel. Others are partially supported by USC Graduate Fellowships, NSF IIS-1451412, 1513966, CCF-1139148 and A. P. Sloan Research Fellowship.
Robust and Accurate Line- and/or Point-Based Pose Estimation without Manhattan Assumptions

Yohann Salaün1,2, Renaud Marlet1, Pascal Monasse1

1LIGM, UMR 8049, École des Ponts, UPE, Champs-sur-Marne, France
2CentraleSupélec, Châtenay-Malabry, France

Motivation

Point matches Line matches

Common failure cases of point-based SfM:
- lack of texture ⇒ few point detections
- wide baseline or little overlap ⇒ few point matches
- planar surfaces ⇒ degenerate cases for computing \( E, F \)

However, line-based SfM not affected.

Two-View Line-Based Pose Estimation

Point matches Line matches, vanishing points

Main existing approach [1] (3-line):
- \( R \) estimated from 3 lines assumed s.t. \( (L_1 \parallel L_2) \perp L_3 \) ⇒ Hyp. \( \exists \perp \) vanishing directions / Hyp. \( \exists \) enough triplets
- \( t \) estimated from line intersections ⇒ poor cues
- \((R, t)\) refined from line intersections ⇒ information loss

Our approach (2x2-line):
- \( R \) estimated from 2 pairs of lines \( (L_1 \parallel L_2), (L_3 \parallel L_4) \) ⇒ no Manhattan-world assumption & \( \exists \) many pairs
- \( t \) estimated from line intersections and/or points
- \((R, t)\) refined from lines and points using angular error

Robust Pose Estimation

- RANSAC-based framework
- sampling of both lines and points (mixed method)
- unifying angular distance for both lines and points

Robust and Parameterless Pose-Estimation

A contrario variant (AC-mixed):
- automatic RANSAC threshold
- single probabilistic framework for both lines & points

\[
NFA(n, k, \epsilon) = 10(n - 6)\binom{n}{k}p(\epsilon)^{k-6}
\]

\[
d_{\text{points}}(p_i, q_i) = \angle(Rp_i \times t, q_i \times t)
\]

\[
d_{\text{line}}(l_i, m_i, R) = \min_{l_j, m_j} d_{\text{line}}(l_i, l_j, m_i, m_j)
\]

\[
p_{\text{line}}(\epsilon) = P(d_{\text{line}}(l_i, m_i, R) \leq \epsilon) = 1 - \cos \epsilon.
\]

Experiments

Lesser sensitivity to noise

Robustness to non-Manhattan configurations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Strecha</td>
<td>R</td>
<td>0.02</td>
<td>0.46</td>
<td>0.25</td>
<td>0.19</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0.18</td>
<td>3.37</td>
<td>1.03</td>
<td>0.80</td>
<td>0.21</td>
</tr>
<tr>
<td>Office</td>
<td>R</td>
<td>6.88</td>
<td>6.45</td>
<td>1.03</td>
<td>1.01</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>27.19</td>
<td>20.38</td>
<td>3.26</td>
<td>3.13</td>
<td>1.44</td>
</tr>
<tr>
<td>Building</td>
<td>R</td>
<td>0.23</td>
<td>6.68</td>
<td>0.49</td>
<td>0.24</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0.31</td>
<td>37.63</td>
<td>1.57</td>
<td>0.83</td>
<td>0.45</td>
</tr>
<tr>
<td>Car</td>
<td>R</td>
<td>0.19</td>
<td>24.25</td>
<td>2.37</td>
<td>0.75</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0.20</td>
<td>69.47</td>
<td>18.03</td>
<td>0.89</td>
<td>0.28</td>
</tr>
</tbody>
</table>

Residual errors: mixed (line-&-point) methods provide extra robustness while preserving accuracy

References

MARLow: A Joint Multiplanar Autoregressive and Low-Rank Approach for Image Completion

Mading Li¹, Jiaying Liu¹, Zhiwei Xiong², Xiaoyan Sun³ and Zongming Guo¹

¹Institution of Computer Science & Technology of Peking University, Beijing, China
²University of Science and Technology of China, Hefei, China
³Microsoft Research Asia, Beijing, China

martinli0822@pku.edu.cn & liujiaying@pku.edu.cn & xzw@mail.ustc.edu.cn & xysun@microsoft.com & guozongming@pku.edu.cn

Abstract

We present a multichannel image completion algorithm called MARLow that combines the low-rank minimization framework and the autoregressive model. MARLow is capable of jointly exploiting the local stationarity across different cross-sections and understanding the correlations among different channels. The experiments show that MARLow has achieved promising performance on both synthetic and real-world datasets.

Introduction

Image completion is a challenging problem that involves understanding the intrinsic content of the given degraded matrices and recovering the missing parts.

1. Notation

- $X$: a degraded matrix
- $Y$: the restored matrix
- $N$: the number of patches
- $L$: the number of channels
- $M$: a low-rank matrix
- $R$: the rank of $M$

2. Problem Formulation

The problem of multichannel image completion can be formulated as:

$$\arg\min_{X, M} \| X - Y \|_F^2 + \lambda \| M \|_* + \mu \| M \|_R$$

where $\lambda$ and $\mu$ are regularization parameters.

3. Problem Transformation

When fixing $M$, the problem turns into a low-rank minimization problem:

$$\arg\min_{X} \| X - Y \|_F^2 + \mu \| X \|_R$$

4. Alternating Minimization

We present an alternating minimization algorithm to solve the problem.

Initialization:

- Fix $M$ and estimate $X$.

For each patch group do:

- Fix $X$, estimate $M$.
- $\text{For } i = 1 \text{ to maxIter do}$
- Patch grouping methods
- $\text{Processing multichannel information separately vs. simultaneously}$

Experimental Results

- Text removal
- Process multichannel information
- Output the restored image.
An Uncertain Future: Forecasting from Variational Autoencoders
Jacob Walker, Carl Doersch, Abhinav Gupta, Martial Hebert
The Robotics Institute, Carnegie Mellon University
Machine Learning Department, Carnegie Mellon University

Goal
Given an image, predict object trajectories.

Prediction 1
1. Utilize VAEs to compute distribution of trajectories conditioned on image.
2. Samples of latent variables represent different future motions in image.
3. Representation learned useful for object detection (especially humans).

Previous Work
Optical Flow
Walker et al., ICCV 2015

Nearest Neighbors
Yuen and Torralba, ECCV 2010

LSTM
Yuan et al., ECCV 2010

Summary
1. Utilize VAEs to compute distribution of trajectories conditioned on image.
2. Samples of latent variables represent different future motions in image.
3. Representation learned useful for object detection (especially humans).

Training Data
Extract Trajectories from Unlabeled Video

Training Architecture
Regress over Trajectories using VAE

Testing Architecture

Multimodal Predictions
Different Samples Give Different Predictions

Quantitative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>NLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressor</td>
<td>11463</td>
</tr>
<tr>
<td>Optical Flow [1]</td>
<td>11734</td>
</tr>
<tr>
<td>Ours</td>
<td>11082</td>
</tr>
</tbody>
</table>

Representation Learning

<table>
<thead>
<tr>
<th>VOC 2012 Test</th>
<th>Data</th>
<th>Person</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch + cal</td>
<td>N/A</td>
<td>51.7</td>
<td>40.6</td>
</tr>
<tr>
<td>Kmeans + cal [2]</td>
<td>N/A</td>
<td>57.1</td>
<td>46.1</td>
</tr>
<tr>
<td>Rel. pos. + cal [3]</td>
<td>1.2M ImageNet</td>
<td>57.5</td>
<td>52.3</td>
</tr>
<tr>
<td>Egomotion [4]</td>
<td>20.5K KITTI</td>
<td>53.5</td>
<td>44.9</td>
</tr>
<tr>
<td>Vid embed [5]</td>
<td>1.5M (100k) videos</td>
<td>53.9</td>
<td>46.2</td>
</tr>
<tr>
<td>Vid embed [5]</td>
<td>5M (100k) videos</td>
<td>54.1</td>
<td>47.0</td>
</tr>
<tr>
<td>Vid embed [5]</td>
<td>8M (100k) videos</td>
<td>54.8</td>
<td>47.5</td>
</tr>
<tr>
<td>Vid embed + cal [5]</td>
<td>8M (100k) videos</td>
<td>48.4</td>
<td>42.2</td>
</tr>
<tr>
<td>Ours + cal</td>
<td>13K UCF101 vid.</td>
<td>58.4</td>
<td>47.3</td>
</tr>
</tbody>
</table>

Latent Space Interpolation

Carried Object Detection based on an Ensemble of Contour Exemplars

Farnoosh Ghadiri\textsuperscript{1}, Robert Bergevin\textsuperscript{1}, Guillaume-Alexandre Bilodeau\textsuperscript{2}

\textsuperscript{1}LVSN-REPARTI, Université Laval, \textsuperscript{2}LITIV lab., Polytechnique Montréal

**Challenge**

People can carry a variety of objects such as handbag, a musical instrument, or even an unusual/dangerous item like an improvised explosive device.

Can we automatically detect all kind of objects carried by people without defining specific model for them?

**Results**

<table>
<thead>
<tr>
<th>Id</th>
<th>View</th>
<th>Position</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**Proposed Method (ECE)**

- Precision: 67%
- Recall: 91%
- TP: 60%
- FP: 21%
- FN: 25%
- F1 Score: 74%

**Damen et al.**

- Precision: 50%
- Recall: 55%
- TP: 46%
- FP: 45%
- FN: 37%
- F1 Score: 52%

**Comparison of Damen et al. & Tavanai et al. with the proposed method over PETS 2006.**

<table>
<thead>
<tr>
<th>Id</th>
<th>View</th>
<th>Position</th>
<th>SC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>142</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>143</td>
<td>8</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

**Conclusion**

- Our experiments indicate that learning human model from human’s contour makes the system more robust to the factors that may give rise to irregularities such as clothing, than methods that model humans based on silhouettes.
- Using biased normalized cut to segment object combined with the high-level information of human model, provides us with a rough estimation of the carried object shape.

**Success & Failure**

- Poor person’s hypothesis
- Due to the inherent variability of pose and cloth appearances.
- Wrongly detected person’s view
- Poor extracted foreground
Determinantal Point Process (DPP):

- DPP is used for subset selection
- Summarization is modeled as subset selection
- Models negative correlations
- Successful in document summarization

Formulation:

- Partition the video into disjoint sets, \( Y_1, \ldots, Y_T \)
- Construct kernel \( L \) for shots in the partition
- Denote by \( Y \) be \( \{1, 2, \ldots, N\} \) the shots in the partition:

\[
P(Y = y) = \det(L_y)/\det(L + I), \quad \forall y \subseteq Y,
\]

- Two layers of DPPs for each partition:
  1) \( Z \)-layer: query-relevant summarization
  2) \( Y \)-layer: contextual summarization
- Condition DPPs through time/hierarchy

Experimental Setup/Results:

| Table 1: Results of query-focused video summarization with two concept queries.

<table>
<thead>
<tr>
<th>User</th>
<th>UDF (%)</th>
<th>TV episode (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sampling</td>
<td>22.12 ( \pm 0.67 )</td>
<td>77.11 ( \pm 24.04 )</td>
</tr>
<tr>
<td>Testing</td>
<td>30.16 ( \pm 3.15 )</td>
<td>77.11 ( \pm 24.04 )</td>
</tr>
<tr>
<td>DPP</td>
<td>13.47</td>
<td>77.11 ( \pm 24.04 )</td>
</tr>
<tr>
<td>SDP</td>
<td>13.47</td>
<td>77.11 ( \pm 24.04 )</td>
</tr>
</tbody>
</table>

Advantages:

- Each layer has its own DPP kernel
- Kernels are learned during training
- Useful for streaming videos

Datasets/Training/Testing:

1) UT Egocentric
2) TV Episodes

- 4 videos
- Each 3~5 hours
- Dense text annotations available for both

Training:

- MLE on joint probability distribution
- Learn linear embedding for each summarization layer

Test:

- Approximate online inference

Experimental Setup:

- Train on 3 videos
- Test on the remaining video

New evaluation metric:

- Hitting recall: ratio of query-relevant shots in the system summary to GT.

Inputs:

- Video
- Query

Output:

- Query-relevant + contextual summary

Applications:

- Obtain video snippets in search engines
- Content-aware video summarization

Some clarifications:

- Input:
  - Video
  - Query

- Query:
  - (a) Input: Video & Query
  - (b) Algorithm: Sequential & Hierarchical Determinantal Point Process (SH-DPP)
  - (c) Output: Summary

Features:

- Sentibank features
- Mean correlation of low level features

Framework

(a) Input: Video & Query
(b) Algorithm: Sequential & Hierarchical Determinantal Point Process (SH-DPP)
(c) Output: Summary

Advantages:

- Each layer has its own DPP kernel
- Kernels are learned during training
- Useful for streaming videos

Applications:

- Obtain video snippets in search engines
- Content-aware video summarization

Videos

- Car
- Children
- Drink
- Flowers
- Street
- Area
- Food
- Water

<table>
<thead>
<tr>
<th>Query1</th>
<th>Summary1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important &amp; diverse shots</td>
<td>Query-relevant, important, &amp; diverse shots</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Query2</th>
<th>Summary2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Important &amp; diverse shots</td>
<td>Query-relevant, important, &amp; diverse shots</td>
</tr>
</tbody>
</table>

Experimental Setup:

- Train on 3 videos
- Test on the remaining video

New evaluation metric:

- Hitting recall: ratio of query-relevant shots in the system summary to GT.

Input:

- Video
- Query

Output:

- Query-relevant + contextual summary

Applications:

- Obtain video snippets in search engines
- Content-aware video summarization

Features:

- Sentibank features
- Mean correlation of low level features

Advantages:

- Each layer has its own DPP kernel
- Kernels are learned during training
- Useful for streaming videos
Temporal Segment Networks: Towards Good Practices for Deep Action Recognition

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool
ETH Zurich The Chinese University of Hong Kong Shenzhen Institutes of Advanced Technology, CAS

Motivation

- Modeling long-range temporal structure is crucial for human activity recognition.
- Frames in a video are highly redundant.

**Modeling long-range temporal structure is not simply wrapping tons of frames. Frames are dense, but contents are sparse!**

Temporal Segment Networks (TSN): The Model

**Training TSN**
1. Divide one video into a fixed number of segments
2. Sample snippets from the segments
3. Optimize the classification loss based on segment consensus

**Temporal Segment Networks (TSN)**

Video | Video Snippets | Temporal Segment Networks
--- | --- | ---
RGB | RGB CNN | RGB CNN
Optical Flow | Optical Flow CNN | Optical Flow CNN

**Segment-Based Sparse Sampling**

$$TSN(T_1, T_2, \ldots, T_K) = \mathcal{H}(\mathcal{G}(\mathcal{F}(T_1; W), \mathcal{F}(T_2; W), \ldots, \mathcal{F}(T_K; W)))$$

**Input Modalities**

- Raw RGB
- Optical Flow
- Warped Flow
- RGB Difference

**Experimental Results**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>HMD851</th>
<th>HMDB51</th>
<th>UCf101</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoFAP Two Stream TDD LTC KVMF TSN</td>
<td>69.40%</td>
<td>93.2%</td>
<td>94.2%</td>
</tr>
</tbody>
</table>

**Component Analysis: Identify the Good Practices**

On UCf101-Split 1, adding components one by one

<table>
<thead>
<tr>
<th>Component</th>
<th>Basic</th>
<th>Two-Stream</th>
<th>Cross-Modality Pre-training</th>
<th>Partial BN with dropout</th>
<th>Temporal Segment Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>90.0%</td>
<td>91.5</td>
<td>92.0%</td>
<td>92.5%</td>
<td>93.5%</td>
</tr>
</tbody>
</table>


Overview

- **Goal**: Determine the geolocation of any photo.
- **Geolocation by classification** using CNNs.
  - Input: photos, output classes: geographical cells.
  - Classification allows multimodal location predictions.
- Localizes landmarks, landscapes, locally typical objects, and even some plants and animals.
- Competitive with Im2GPS, outperforms humans.

Geographical Partitioning Scheme

- Adaptive partitioning using S2 cells.
  - Recursive splitting until no cell contains over $t_1$ photos.
  - Discard cells with less than $t_2$ photos.
  - Even class distribution, adaptive spatial resolution.
  - Up to street level resolution in densely populated areas.

Training

- **Data**: 91M photos from the web with Exif locations.
- 26,263 classes ($t_1=10,000$, $t_2=50$). Target is 1-hot vector.
- Inception architecture, 97M parameters, trained 2.5 months.

Results

- Correctly localized photos
- Incorrectly localized photos
- Top photos by region

PlaNet vs. Im2GPS

**Im2GPS**: Localizes query by nearest neighbor matching against Flickr photos. Im2GPS (new) adds geo-clustering, 1-vs-all SVM.

<table>
<thead>
<tr>
<th>Method</th>
<th>Index / Training Images</th>
<th>Street 1km</th>
<th>City 25km</th>
<th>Region 200km</th>
<th>Country 750km</th>
<th>Continent 250km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Im2GPS (orig)</td>
<td>6.5M</td>
<td>12.0%</td>
<td>15.0%</td>
<td>23.0%</td>
<td>47.0%</td>
<td></td>
</tr>
<tr>
<td>Im2GPS (new)</td>
<td>6.5M</td>
<td>2.5%</td>
<td>21.9%</td>
<td>32.1%</td>
<td>35.4%</td>
<td>51.9%</td>
</tr>
<tr>
<td>PlaNet</td>
<td>900k</td>
<td>0.4%</td>
<td>3.8%</td>
<td>7.6%</td>
<td>21.6%</td>
<td>43.5%</td>
</tr>
<tr>
<td>PlaNet</td>
<td>6.2M</td>
<td>6.3%</td>
<td>18.1%</td>
<td>30.0%</td>
<td>45.6%</td>
<td>65.8%</td>
</tr>
<tr>
<td>PlaNet</td>
<td>91M</td>
<td>8.4%</td>
<td>24.5%</td>
<td>37.6%</td>
<td>53.6%</td>
<td>71.3%</td>
</tr>
</tbody>
</table>

Model comparison

- Features: Trained jointly and end-to-end
- Prediction: NN matching + geo-clustering + 1-vs-all SVM
- Size: 377 MB vs. 577 GB

Playing GeoGuessr against Humans

- **Goal**: Guess the location of a random street view panorama.
- 10 humans played 50 rounds against PlaNet. PlaNet won 28.
- Median error: 1131.7km (PlaNet), 2320.8km (humans)

Model Analysis

Comparison to retrieval

Reducing model size

Geolocating Photo Albums

- **Idea**: Context can help disambiguate hard-to-localize photos.
  - Using LSTM (long short term memory) architecture to predict the sequence of geolocations in photo albums.
  - Data: 29.7M albums (616M photos), split 80% / 20%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Street 1km</th>
<th>City 25km</th>
<th>Region 200km</th>
<th>Country 750km</th>
<th>Continent 250km</th>
</tr>
</thead>
<tbody>
<tr>
<td>PlaNet</td>
<td>14.9%</td>
<td>20.3%</td>
<td>27.4%</td>
<td>42.0%</td>
<td>61.8%</td>
</tr>
<tr>
<td>PlaNet avg</td>
<td>22.2%</td>
<td>35.6%</td>
<td>51.4%</td>
<td>58.6%</td>
<td>82.7%</td>
</tr>
<tr>
<td>PlaNet HMM</td>
<td>23.3%</td>
<td>34.3%</td>
<td>47.1%</td>
<td>63.2%</td>
<td>79.5%</td>
</tr>
<tr>
<td>LSTM</td>
<td>32.0%</td>
<td>42.1%</td>
<td>57.9%</td>
<td>75.5%</td>
<td>87.9%</td>
</tr>
<tr>
<td>LSTM off</td>
<td>30.9%</td>
<td>41.0%</td>
<td>56.9%</td>
<td>74.3%</td>
<td>85.4%</td>
</tr>
<tr>
<td>LSTM of2</td>
<td>29.9%</td>
<td>40.0%</td>
<td>55.6%</td>
<td>73.4%</td>
<td>83.5%</td>
</tr>
<tr>
<td>LSTM rep</td>
<td>34.5%</td>
<td>45.6%</td>
<td>62.6%</td>
<td>79.3%</td>
<td>90.5%</td>
</tr>
<tr>
<td>LSTM rep 25</td>
<td>28.3%</td>
<td>37.5%</td>
<td>49.9%</td>
<td>68.9%</td>
<td>82.0%</td>
</tr>
<tr>
<td>BLSTM 25</td>
<td>33.0%</td>
<td>43.0%</td>
<td>56.7%</td>
<td>73.2%</td>
<td>86.1%</td>
</tr>
</tbody>
</table>
Detecting Text in Natural Image with Connectionist Text Proposal Network

Zhi Tian¹, Weilin Huang¹,², Tong He¹, Pan He¹, and Yu Qiao¹,³
¹ Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
² University of Oxford, UK ³ The Chinese University of Hongkong, China

Online demo: textdet.com

**Motivation**
- Current bottom-up approaches are complicated, with weak robustness and reliability, and accumulated errors.
- Stat-of-the-art object detectors are powerful, but not accurate for text localisation.
- Fill the gap between general object detection (e.g., RPN [1]) and text detection.

**Connectionist Text Proposal Network**

**CTPN Architecture**
- Detect text in sequences of fine-scale proposals
- Recurrently connect sequential proposals by BLSTM
- Jointly predict text scores, y-axis coordinates, and refinement offsets

**CTPN Proposals**

**Detecting Text in Fine-Scale Proposals**

**Experimental Results**

**Recurrent Connectionist Text Proposals**

**Top:** CTPN without recurrent connection. **Bottom:** with recurrent connection
- RNN layer connects sequential proposals directly in convolutional layer
- In-network recurrent architecture is end-to-end trainable
- Encode rich context information
- Detect highly ambiguous text, and reduce false detections considerably

**Side-Refinement**
- Predict offsets for side-proposals - horizontal sides rectification
- Further improve localisation accuracy
- Joint predictions - not a post-precessing step

Reference:

**Table 1:** Component evaluation on the ICDAR 2013, and State-of-the-art results on the SWT and MULTILINGUAL.

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPN</td>
<td>0.17</td>
<td>0.63</td>
<td>0.27</td>
<td>RPN (no RNN)</td>
<td>0.83</td>
<td>0.79</td>
<td>0.80</td>
<td>CTPN</td>
<td>0.68</td>
<td>0.60</td>
<td>0.66</td>
<td>CTPN</td>
<td>0.83</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>0.66</td>
<td>0.69</td>
<td>0.69</td>
<td>Faster R-CNN</td>
<td>0.66</td>
<td>0.69</td>
<td>0.69</td>
<td>Faster R-CNN</td>
<td>0.66</td>
<td>0.69</td>
<td>0.69</td>
<td>CTPN</td>
<td>0.68</td>
<td>0.60</td>
<td>0.66</td>
</tr>
<tr>
<td>Text-Net [18]</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>Text-Net</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>Text-Net</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>CTPN</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>CNN-Text [19]</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>CNN-Text</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>CNN-Text</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>CTPN</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
</tr>
</tbody>
</table>

**Table 2:** State-of-the-art results on the ICDAR 2011, 2013 and 2015.

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huang [3]</td>
<td>0.82</td>
<td>0.78</td>
<td>0.60</td>
<td>Huang [3]</td>
<td>0.84</td>
<td>0.73</td>
<td>0.76</td>
<td>Huang [3]</td>
<td>0.84</td>
<td>0.73</td>
<td>0.76</td>
<td>Huang [3]</td>
<td>0.84</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td>You [11]</td>
<td>0.84</td>
<td>0.80</td>
<td>0.80</td>
<td>You [11]</td>
<td>0.85</td>
<td>0.80</td>
<td>0.80</td>
<td>You [11]</td>
<td>0.85</td>
<td>0.80</td>
<td>0.80</td>
<td>You [11]</td>
<td>0.85</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Yin [14]</td>
<td>0.85</td>
<td>0.80</td>
<td>0.80</td>
<td>Yin [14]</td>
<td>0.86</td>
<td>0.80</td>
<td>0.80</td>
<td>Yin [14]</td>
<td>0.86</td>
<td>0.80</td>
<td>0.80</td>
<td>Yin [14]</td>
<td>0.86</td>
<td>0.80</td>
<td>0.80</td>
</tr>
<tr>
<td>Zhang [34]</td>
<td>0.84</td>
<td>0.78</td>
<td>0.60</td>
<td>Zhang [34]</td>
<td>0.84</td>
<td>0.78</td>
<td>0.60</td>
<td>Zhang [34]</td>
<td>0.84</td>
<td>0.78</td>
<td>0.60</td>
<td>Zhang [34]</td>
<td>0.84</td>
<td>0.78</td>
<td>0.60</td>
</tr>
<tr>
<td>Text-Net [29]</td>
<td>0.86</td>
<td>0.76</td>
<td>0.68</td>
<td>Text-Net [29]</td>
<td>0.86</td>
<td>0.76</td>
<td>0.68</td>
<td>Text-Net [29]</td>
<td>0.86</td>
<td>0.76</td>
<td>0.68</td>
<td>Text-Net [29]</td>
<td>0.86</td>
<td>0.76</td>
<td>0.68</td>
</tr>
<tr>
<td>CNN-Text [11]</td>
<td>0.90</td>
<td>0.87</td>
<td>0.84</td>
<td>CNN-Text [11]</td>
<td>0.90</td>
<td>0.87</td>
<td>0.84</td>
<td>CNN-Text [11]</td>
<td>0.90</td>
<td>0.87</td>
<td>0.84</td>
<td>CNN-Text [11]</td>
<td>0.90</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>CTPN</td>
<td>0.89</td>
<td>0.80</td>
<td>0.80</td>
<td>CTPN</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>CTPN</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>CTPN</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
</tr>
</tbody>
</table>

**Summary:**
- Trained on 3K images in English and Chinese, generalise well to others (e.g., Korean)
- Fine-scale strategy improves Precision, while using RNN increases Recall and Precision
- Obtain 0.88 and 0.61 F-measures on the ICDAR 2013 and 2015, respectively
- Computationally efficient, with 0.14s/image GPU time (scale=600)
- Strong capability for detecting very small-size text

**Experimental Results**

**Red Box:** CTPN detection. **Yellow Box:** ground truth

**Reference:**
Face recognition using a unified 3D morphable model

G Hu, F Yan, C Chan, W Deng, W Christmas, J Kittler, N Robertson
Anyvision, University of Surrey, BUPT, Queen’s University Belfast
huguosheng100@gmail.com

Motivation

3D morphable model (3DMM)

$3_0 = 3_3 + T3$

Problems

- Limited modelling capacity
- Pose, illumination and expression only
- Fitting difficulties
- Inaccurate, slow
- Expensive 3D training data

Unified 3D Morphable Model (U-3DMM) Modelling

$3_0 = 3_3 + T3$

Inter-personal variation

Extra-personal variation

Extra-personal variation

Pose, occlusion and illumination

Shape and texture parameters are concatenated as face representation. U-3DMM is compared with Traditional 3DMM, Extended 3DMM, Sparse Representation Classification (SRC), Extended SRC (ESRC), deep learning methods. We used Multi-PIE, AR, and a synthetic database to evaluate our method.

Fitting

Intra-personal variation data collection

Experiments

Settings

Pose and illumination

Table 4: Recognition rate (%) averaging 20 illuminations on Multi-PIE

Table 5: Recognition rate (%) evaluated on AR database

$3_0 = 3_3 + T3$

3D Shape Model

$3_0 = 3_3 + T3$

Conclusions

- Strong modeling capacity
- Accurate and efficient fitting
- 2D images for training

Anyvision interactive technologies Ltd.
Augmented Feedback in Semantic Segmentation under Image Level Supervision

Xiaojuan Qi¹, Zhengzhe Liu¹, Jianping Shi², Hengshuang Zhao¹, Jiaya Jia¹
¹The Chinese University of Hong Kong
²SenseTime Group Limited

Introduction

Motivation

• Hand labeling every pixel for semantic segmentation is labor intensive. (Training data, Scalability).
• Learn to segment semantic objects with image level labels.

Previous work

• Multiple learning based method[1][2].
• Expectation-Maximization based method[3].
  E-step: Estimate the interim label with the learned model
  M-step: Train the model with the interim label as ground truth.
  Easy to be locked in local minimum.

Overall Algorithm

Algorithm 2: Our Complete Network System with Augmented Feedback

Input:
  Input image set; proposal set; maximum iteration number T;
Procedure:
  1: Initialize the proposal score with the classification network [32, 35];
  2: for i ∈ [1, T] do
  3:  Aggregate proposals to generate segmentation masks (described in Sect. 4.1.1);
  4:  Train semantic segmentation branch;
  5:  Select positive and negative samples (described in Sect. 4.2.1);
  6:  Train object localization branch;
  7:  Back-annotate object proposals using the trained localization branch;
  end for;
Output: Semantic segmentation model.

Our Method – Augmented feedback loop for segmentation and localization

Motivation

• Hand labeling every pixel for semantic segmentation is labor intensive. (Training data, Scalability).
• Learn to segment semantic objects with image level labels.

Previous work

• Multiple learning based method[1][2].
• Expectation-Maximization based method[3].
  E-step: Estimate the interim label with the learned model
  M-step: Train the model with the interim label as ground truth.
  Easy to be locked in local minimum.

Overall Algorithm

Algorithm 2: Our Complete Network System with Augmented Feedback

Input:
  Input image set; proposal set; maximum iteration number T;
Procedure:
  1: Initialize the proposal score with the classification network [32, 35];
  2: for i ∈ [1, T] do
  3:  Aggregate proposals to generate segmentation masks (described in Sect. 4.1.1);
  4:  Train semantic segmentation branch;
  5:  Select positive and negative samples (described in Sect. 4.2.1);
  6:  Train object localization branch;
  7:  Back-annotate object proposals using the trained localization branch;
  end for;
Output: Semantic segmentation model.

Results

• We achieve 55.5% mean IoU on Pascal voc 2012 test set compared to 40.6%.
• The localization result has been improved from 5.71% mAP to 28.25% on the voc validation set.
• The segmentation result has been improved from 38.63% to 54.34% on validation set.
Towards Viewpoint Invariant 3D Human Pose Estimation

Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, Li Fei-Fei

Introduction

- Depth cameras are becoming cheaper and more ubiquitous
- In most real-world applications, cameras are deployed with challenging viewpoints (e.g., ceiling of retail stores, angled views at airports)
- These viewpoints are difficult for existing pose estimation algorithms

**Input:** Single depth image  **Output:** Body part locations in 3D

We propose a viewpoint invariant model for human pose estimation that embeds local regions onto a learned feature space with a multi-task loss.

Towards Viewpoint Invariance

Input Representation

A glimpse is a retina-like encoding of the input [1]. Pixels further from the glimpse center are blurred.

Two different glimpses:

- Glimpses force our model to focus on specific regions instead of analyzing the full input image
- We use one glimpse per body part

Learned Viewpoint Invariant Feature Space

1. Input glimpses are converted to a point cloud
2. A spatial transformer [2] warps the cloud and can create artificial occlusions that may occur in other viewpoints
3. After training, the warped point cloud can be interpreted as a viewpoint invariant feature embedding of 3D body parts

Method

**Our method:**

- Is formulated as a multi-task learning problem to selectively predict partial poses in case of noise or occlusion
- Learns geometric body part variations
- Combines a CNN and RNN to self-correct previous pose estimates

Multi-Task Loss

Partial Poses

- In extreme viewpoints (e.g., top view), many body parts are occluded
- To account for this, our model simultaneously predicts:
  - The location of each body part (real values)
  - Whether the body part is visible or occluded (binary)

Datasets

- Existing depth-based human pose datasets are small in size, both in number of images and number of classes
- We collected a new dataset of 100K depth images with pixel-wise body part labels and 3D human body part locations

Stanford EVAL Dataset [5]  Invariant-Top View Dataset (Ours)

- The dataset has been made publicly available online at:
  
References

Motivation

The Object Target Representation Model

- **Global Model** is effective to holistic appearance changes, like illumination variations and pose changes.
- **Local Model** is intrinsically robust to the challenges, such as partial occlusions and local deformations.

**Goal of this work**

To leverage the effectiveness of global model in capturing overall information, and augment it with a local method to promote the robustness of the tracker.

Solution

**Basic Idea**: according to the target summarization (global) and the target priors (local) to estimate an expected target.

\[
\hat{y}_k = \varphi(y_1, y_2, \ldots, y_{k-1} | \Phi)
\]

The candidate \(c\) the most similar to the expected target is determined as the target.

\[
y_k = \arg\min_{c \in C} \| \hat{y}_k - c \|
\]

Subspace Model summarizes the temporal targets. Local Observation offers priors of the target. Matrix Completion estimates the expected target and maintains the subspace structure.

Experimental Results

Performance on our dataset containing 20 sequences

Low dimension verification

Local observation verification

In the case of deformation

In the case of occlusion

Visualization of the local observation
In this work, we present the Inter-Battery Topic Model (IBTM). Our approach extends traditional topic models by learning a factorized latent variable representation. The structured representation leads to a model that marries benefits traditionally associated with a discriminative approach, such as feature selection, with those of a generative model, such as principled regularization and ability to handle missing data. The factorization is provided by representing data in terms of aligned pairs of observations as different views. This provides means for selecting a representation that separately models topics that exist in both views from the topics that are unique to a single view. This structured association allows for efficient and robust inference and provides a compact and efficient representation. Learning is performed in a Bayesian fashion by maximizing a rigorous bound on the log-likelihood. The model is then evaluated in both uni- and multi-modality settings on two different classification tasks with off-the-shelf convolutional neural network (CNN) features that generate state-of-the-art results with extremely compact representations. Additionally, we introduce a novel application with IBTM in healthcare.

Inter-Battery Topic Representation Learning

Inter-Battery Topic Representation Learning

INTER-BATTERY TOPIC REPRESENTATION LEARNING
Cheng Zhang, Hedvig Kjellström and Carl Henrik Ek
{chengz, hedvig}@kth.se, carlhenrik.ek@bristol.ac.uk

ABSTRACT

In this work, we present the Inter-Battery Topic Model (IBTM). Our approach extends traditional topic models by learning a factorized latent variable representation. The structured representation leads to a model that marries benefits traditionally associated with a discriminative approach, such as feature selection, with those of a generative model, such as principled regularization and ability to handle missing data. The factorization is provided by representing data in terms of aligned pairs of observations as different views. This provides means for selecting a representation that separately models topics that exist in both views from the topics that are unique to a single view. This structured association allows for efficient and robust inference and provides a compact and efficient representation. Learning is performed in a Bayesian fashion by maximizing a rigorous bound on the log-likelihood. The model is then evaluated in both uni- and multi-modality settings on two different classification tasks with off-the-shelf convolutional neural network (CNN) features that generate state-of-the-art results with extremely compact representations. Additionally, we introduce a novel application with IBTM in healthcare.

INTRODUCTION

In this work, we present the Inter-Battery Topic Model (IBTM). Our approach extends traditional topic models by learning a factorized latent variable representation. The structured representation leads to a model that marries benefits traditionally associated with a discriminative approach, such as feature selection, with those of a generative model, such as principled regularization and ability to handle missing data. The factorization is provided by representing data in terms of aligned pairs of observations as different views. This provides means for selecting a representation that separately models topics that exist in both views from the topics that are unique to a single view. This structured association allows for efficient and robust inference and provides a compact and efficient representation. Learning is performed in a Bayesian fashion by maximizing a rigorous bound on the log-likelihood. The model is then evaluated in both uni- and multi-modality settings on two different classification tasks with off-the-shelf convolutional neural network (CNN) features that generate state-of-the-art results with extremely compact representations. Additionally, we introduce a novel application with IBTM in healthcare.

MODEL

The figure above visualizes the shared topic representation (θ) and private topic representations (α and ν). The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters.

EXPERIMENT

In this experiment, we explore the scenario in which only one modality is available. Both views use bag-of-Conv1 features and are represented as images in terms of aligned pairs of observations as different views. This provides means for selecting a representation that separately models topics that exist in both views from the topics that are unique to a single view. This structured association allows for efficient and robust inference and provides a compact and efficient representation. Learning is performed in a Bayesian fashion by maximizing a rigorous bound on the log-likelihood. The model is then evaluated in both uni- and multi-modality settings on two different classification tasks with off-the-shelf convolutional neural network (CNN) features that generate state-of-the-art results with extremely compact representations. Additionally, we introduce a novel application with IBTM in healthcare.

Application on diagnostic Prediction

The goal of this application [4] is to predict possible diagnostic results given a discomfort drawing image. The figure above visualizes the shared topic representation (θ) and private topic representations (α and ν). The documents of different classes are colored differently and the plots show the first three principal components after applying PCA on the per document topic distributions for all the training data. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters. The figure on the left shows the histogram over partition parameters.

Acknowledgement

This research has been supported by the Swedish Research Council (VR), Stiftelsen Promobilia.
Online Adaptation for Joint Scene and Object Classification

Jawadul H. Bappy, Sujoy Paul, and Amit K. Roy-Chowdhury

Introduction
- Most existing methods assume that data will be labeled and available beforehand to train the classification models.
- It becomes infeasible and unrealistic to know all the labels beforehand with the huge corpus of visual data being generated on a daily basis.
- Adaptability of the models to the incoming data is crucial too for long-term performance guarantees.

Question!!
- Are all the samples equally important to manually label and learn a model?

Motivation
- Scene and objects co-occur in an image. Existing active learning methods focus on individual classification task (either scene or object).
- Can we exploit scene & object inter-relationships to select the most informative samples in order to classify them?

Main Contributions
- The proposed active learning framework learns scene and object classification models simultaneously.
- Both scene and object classification models take advantage of the interdependence between them to select the most informative samples with the least manual labeling cost. To the best of our knowledge, any previous work using active learning to classify scene and objects together is unknown.
- Leveraging upon the inter-relationships between scene and objects, we propose a new information-theoretic sample selection strategy.

Framework Overview
- Scene and object samples are represented as nodes of a graph, and edges delineate the inter-relationship between the samples.
- Selection of an image (or graph) that contains the most informative samples (scene, objects).
- Given an image, a sample (i.e., a node in the graph) is chosen in a way that reduces the uncertainty on other samples.

Experimental Results
- Proposed framework has been evaluated on four datasets MSRC, SUN, MIT-67 Indoor datasets.

Mathematical Formulation.

Scene and Object Model
- Scene Classification: CNN feature, SVM Classifier
- Object Detection: R-CNN

Graphical Model Representation

Formulation of Joint Entropy of a Graph

Subset Selection of Samples

Online Learning for Sample Selection

Summary
- A novel active learning framework for joint scene and object classification is presented by exploiting the interrelationships between them.
- In this paper, we show that both scene and objects interact each other to select the most informative samples in order to learn both of the recognition models.
- Our approach significantly reduces the human effort in labeling samples.

Acknowledgement
This work was supported by NSF grant IIS-1316934 and US ONR contract N00014-15-C-3113 through Mayachitra, inc.
We introduce the concept of unconstrained real-time 3D facial performance capture through explicit semantic segmentation in the RGB input. To ensure robustness, cutting edge supervised learning approaches rely on large training datasets of face images captured in the wild. While impressive tracking quality has been demonstrated for faces that are largely visible, any occlusion due to hair, accessories, or hand-to-face gestures would result in significant visual artifacts and loss of tracking accuracy. The modeling of occlusions has been mostly avoided due to its immense space of appearance variability. To address this curse of high dimensionality, we perform tracking in unconstrained images assuming non-face regions can be fully masked out. Along with recent breakthroughs in deep learning, we demonstrate that pixel-level facial segmentation is possible in real-time by repurposing convolutional neural networks designed originally for general semantic segmentation. We develop an efficient architecture based on a two-stream deconvolution network with complementary characteristics, and introduce carefully designed training samples and data augmentation strategies for improved segmentation accuracy and robustness. We adopt a state-of-the-art regression-based facial tracking framework with segmented face images as training, and demonstrate accurate and uninterrupted facial performance capture in the presence of extreme occlusion and even side views. Furthermore, the resulting segmentation can be directly used to composite partial 3D face models on the input images and enable seamless facial manipulation tasks, such as virtual make-up or face replacement.
Learning Temporal Transformations From Time-Lapse Videos
Yipin Zhou and Tamara L. Berg

Based on life-long observations of the natural world, humans can often easily picture in their minds what an object will look like during a state transformation. What about computers?

Overview

- Learn computational models of object transformations from time-lapse videos.
- In particular, we explore the use of generative models on three tasks to create depictions of objects at future times.

Dataset

Timelapse videos show an entire transformation within a short period of time. (1477 videos from YouTube.)

Example frames from the dataset

Generation tasks

Pairwise generation

Two-stack generation

Recurrent generation

Training & Loss Function

Supplementary video

Experiment results

<table>
<thead>
<tr>
<th>Pairwise</th>
<th>Two stack</th>
<th>Recurrent</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>SSIM</td>
<td>MSE</td>
</tr>
<tr>
<td>p.mse+adv</td>
<td>17.0409</td>
<td>0.57079</td>
</tr>
<tr>
<td>p.g.mse+adv</td>
<td>17.0878</td>
<td>0.5721</td>
</tr>
<tr>
<td>p.g.mse+adv-ft</td>
<td>17.4794</td>
<td>0.6035</td>
</tr>
</tbody>
</table>

Human evaluation

<table>
<thead>
<tr>
<th>Pairwise</th>
<th>Two stack</th>
<th>Recurrent</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>ADV Grad</td>
<td>Ft</td>
</tr>
<tr>
<td>Blooming</td>
<td>0.1320</td>
<td>0.2000</td>
</tr>
<tr>
<td>Melting</td>
<td>0.1680</td>
<td>0.2520</td>
</tr>
<tr>
<td>Baking</td>
<td>0.1620</td>
<td>0.2600</td>
</tr>
<tr>
<td>Rotting</td>
<td>0.1340</td>
<td>0.2020</td>
</tr>
<tr>
<td>Average</td>
<td>0.1490</td>
<td>0.2360</td>
</tr>
</tbody>
</table>
Key Contributions:
- Interactive image segmentation algorithm that
  - Deals with both scribble-based and boundary-based input modalities.
  - Asks the user to provide only foreground pixels (in the case of noiseless scribble inputs).
  - Is robust in the presence of input noise.

Dominant Sets
- Edge-weighted generalization of maximal cliques [1,2]
- Let $G = (V,E,w)$ be edge-weighted graph, and $A$ be its weighted adjacency matrix. Consider,
  $\text{maximize } f(x) = x^T Ax$ subject to $x \in \Delta$
  $\Delta = \{x \in \mathbb{R}^n : e^T x = 1, \text{ and } x \geq 0\}$
- If $x$ is a local maximizer, its support, $\sigma(x)$, is a dominant set

Constrained Dominant Sets
- Regularized version of dominant sets framework
- Given $S \subset V$ and let $I_S$ be a diagonal matrix whose $i^{th}$ diagonal element $d_i = 1$ if $i \in V \setminus S$ and zero otherwise. Consider,
  $\text{maximize } f^\alpha_S(x) = x^T (A - \alpha I_S) x$
  subject to $x \in \Delta$

**Theorem:** Let $S \subset V$ with $S \neq \emptyset$ and let $\alpha > \lambda_{\text{max}}(A|_{V \setminus S})$.
- If $x$ is a local maximizer of $f^\alpha_S$ in $\Delta$, then $\sigma(x) \cap S \neq \emptyset$

Examples

Experimental Results
- Scribble Based Segmentation
- Error-tollerant Scribble Based Segmentation
- Bounding-box approaches with different levels of looseness

Acknowledgement: This work has been partly supported by Samsung GRO.

References:
Deep Markov Random Field for Image Modeling  
Zhirong Wu, Dahua Lin, Xiaoou Tang  
The Chinese University of Hong Kong

http://github.com/zhirongw/deep-mrf

Motivations
- While MRF provides a generic framework for modeling images, the edge potential is largely limited by simplistic functions.
- We wish to improve the expressive power by modeling the inter-pixel relations via deep factors.

Overview
We consider each pixel $x$ is associated with a hidden state $h$.

Deep MRF Formulation
Joint distribution:
$$p(x, h) = \frac{1}{Z} \prod_{u \in V} \zeta(x_u, h_u) \prod_{(u,v) \in E} (\phi(h_u, h_v) \psi(h_u, x_v) \psi(h_v, x_u)) \prod_{u \in V} \lambda(h_u)$$

The potential factors:
- **GMM pixel generation**
  $$\zeta(x_u, h_u) = \mathcal{P} \mathcal{G}(x_u | h_u)$$
- **nearby pixel and state**
  $$\phi(h_u, h_v) = \exp(h_u^T W h_v)$$
- **states regularizer**
  $$\psi(h_u, x_v) = \exp(h_u^T R x_v)$$
- **state regularizer**
  $$\lambda(h_u) = \exp(-1^T \eta(h_u))$$

Connections with RNN
MAP inference of hidden states corresponds to feed-forward computation of an RNN,
$$\tilde{h}_u = \sigma \left( \sum_{v \in N_u} W h_v + R x_v \right)$$

The activation function of RNN is derived from the regularizer of MRF,
$$\sigma^{-1}(z) = \eta'(z)$$

Example activations:
- ReLU
- Regularizer
- Sigmoid
- Regularizer

Connections with CNN
Performance against CNN

<table>
<thead>
<tr>
<th>Set</th>
<th>PSNR</th>
<th>CNN</th>
<th>SCN</th>
<th>Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 5</td>
<td>32.30</td>
<td>33.10</td>
<td>33.14</td>
<td>33.14</td>
</tr>
<tr>
<td>Set 14</td>
<td>29.00</td>
<td>29.41</td>
<td>29.38</td>
<td>29.38</td>
</tr>
<tr>
<td>BSD</td>
<td>28.20</td>
<td>28.50</td>
<td>28.54</td>
<td>28.54</td>
</tr>
</tbody>
</table>

Image Super Resolution
Add another connection from the low-res pixel.

Image Generation
Combines with VAE for global structure.

Texture Synthesis
High resolution visually realistic textures!

Learning via Coupled Passes
Decouple a cyclic graph into multiple acylic passes while maintaining full contextual reasoning.

Connections with FRAME
- presents a new powerful MRF model
- theoretical connections between MRF and RNN
- nice results on a variety of low-level applications

Connections with FoE
- patch-based filters
- patch-based filters

Deep MRF Formulation
- FRAME: handcrafted
- Neural Networks
- FoE: patch-based filters

Deep MRF Formulation
- Joint distribution:
  $$p(x, h) = \frac{1}{Z} \prod_{u \in V} \zeta(x_u, h_u) \prod_{(u,v) \in E} (\phi(h_u, h_v) \psi(h_u, x_v) \psi(h_v, x_u)) \prod_{u \in V} \lambda(h_u)$$

The potential factors:
- **GMM pixel generation**
  $$\zeta(x_u, h_u) = \mathcal{P} \mathcal{G}(x_u | h_u)$$
- **nearby pixel and state**
  $$\phi(h_u, h_v) = \exp(h_u^T W h_v)$$
- **states regularizer**
  $$\psi(h_u, x_v) = \exp(h_u^T R x_v)$$
- **state regularizer**
  $$\lambda(h_u) = \exp(-1^T \eta(h_u))$$

Connections with RNN
MAP inference of hidden states corresponds to feed-forward computation of an RNN,
$$\tilde{h}_u = \sigma \left( \sum_{v \in N_u} W h_v + R x_v \right)$$

The activation function of RNN is derived from the regularizer of MRF,
$$\sigma^{-1}(z) = \eta'(z)$$

Example activations:
- ReLU
- Regularizer
- Sigmoid
- Regularizer

Connections with CNN
Performance against CNN

<table>
<thead>
<tr>
<th>Set</th>
<th>PSNR</th>
<th>CNN</th>
<th>SCN</th>
<th>Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 5</td>
<td>32.30</td>
<td>33.10</td>
<td>33.14</td>
<td>33.14</td>
</tr>
<tr>
<td>Set 14</td>
<td>29.00</td>
<td>29.41</td>
<td>29.38</td>
<td>29.38</td>
</tr>
<tr>
<td>BSD</td>
<td>28.20</td>
<td>28.50</td>
<td>28.54</td>
<td>28.54</td>
</tr>
</tbody>
</table>

Image Super Resolution
Add another connection from the low-res pixel.

Image Generation
Combines with VAE for global structure.

Texture Synthesis
High resolution visually realistic textures!

Learning via Coupled Passes
Decouple a cyclic graph into multiple acylic passes while maintaining full contextual reasoning.

Connections with FRAME
- presents a new powerful MRF model
- theoretical connections between MRF and RNN
- nice results on a variety of low-level applications

Connections with FoE
- patch-based filters
- patch-based filters

Deep MRF Formulation
- Joint distribution:
  $$p(x, h) = \frac{1}{Z} \prod_{u \in V} \zeta(x_u, h_u) \prod_{(u,v) \in E} (\phi(h_u, h_v) \psi(h_u, x_v) \psi(h_v, x_u)) \prod_{u \in V} \lambda(h_u)$$

The potential factors:
- **GMM pixel generation**
  $$\zeta(x_u, h_u) = \mathcal{P} \mathcal{G}(x_u | h_u)$$
- **nearby pixel and state**
  $$\phi(h_u, h_v) = \exp(h_u^T W h_v)$$
- **states regularizer**
  $$\psi(h_u, x_v) = \exp(h_u^T R x_v)$$
- **state regularizer**
  $$\lambda(h_u) = \exp(-1^T \eta(h_u))$$

Connections with RNN
MAP inference of hidden states corresponds to feed-forward computation of an RNN,
$$\tilde{h}_u = \sigma \left( \sum_{v \in N_u} W h_v + R x_v \right)$$

The activation function of RNN is derived from the regularizer of MRF,
$$\sigma^{-1}(z) = \eta'(z)$$

Example activations:
- ReLU
- Regularizer
- Sigmoid
- Regularizer

Connections with CNN
Performance against CNN

<table>
<thead>
<tr>
<th>Set</th>
<th>PSNR</th>
<th>CNN</th>
<th>SCN</th>
<th>Our</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 5</td>
<td>32.30</td>
<td>33.10</td>
<td>33.14</td>
<td>33.14</td>
</tr>
<tr>
<td>Set 14</td>
<td>29.00</td>
<td>29.41</td>
<td>29.38</td>
<td>29.38</td>
</tr>
<tr>
<td>BSD</td>
<td>28.20</td>
<td>28.50</td>
<td>28.54</td>
<td>28.54</td>
</tr>
</tbody>
</table>

Image Super Resolution
Add another connection from the low-res pixel.

Image Generation
Combines with VAE for global structure.
Contributions

- A novel prior for surface reconstruction which favors symmetric shapes
- Explicit modeling of unexplored areas to detect and enforce symmetries
- Support for arbitrary symmetry types with arbitrary local support domain
- Support for multiple symmetries with overlapping domains allowing to compose symmetries with complex group structure
- Little runtime overhead during surface optimization

Motivation

Many object exhibit symmetries that can be leveraged to denoise, complete, or hallucinate surface parts during surface reconstruction.

Variational Approach

Symmetry distance [Kazhdan et al. 2003]

$$SD(u, \gamma) = \|u - \Pi_\gamma(u)\| = \frac{1}{2}\|u - \gamma(u)\|$$

Idea: Minimize symmetry distance during surface optimization.

$$E(u) = \int \left( |Du| + \lambda f(u) \right) dx + \mu \sum_{\gamma \in \Gamma} \omega_{\gamma} \cdot SD_{\gamma}(u, \gamma)$$

Symmetry composition:

- symmetry 1
- symmetry 2
- symmetry 1+2

Results - Hallucination

Symmetries are detected via a Hough space voting or a RANSAC approach.

Results - Denoising

Variational Approach

Symmetry distance [Kazhdan et al. 2003]

$$SD(u, \gamma) = \|u - \Pi_\gamma(u)\| = \frac{1}{2}\|u - \gamma(u)\|$$

Idea: Minimize symmetry distance during surface optimization.

$$E(u) = \int \left( |Du| + \lambda f(u) \right) dx + \mu \sum_{\gamma \in \Gamma} \omega_{\gamma} \cdot SD_{\gamma}(u, \gamma)$$

Symmetry composition:

- symmetry 1
- symmetry 2
- symmetry 1+2

Results - Prior Strength

raw input + symmetry plane

without symmetry ($\mu = 0$)

with symmetry ($\mu = 2\times$)
SPLeaP: Soft Pooling of Learned Parts for Image Classification

Introduction

Part-Based Models (PBM) in the context of Image Classification.

Objective:
Learn \((\Theta, W)\) using annotated training dataset \(D = \{(x_n, y_n)\}_{n=1}^{N}\), with \(x_n = (x_{n1}, \ldots, x_{ni})\) and \(y_n \in \{1, \ldots, C\}\).

Region Description:
\(X = (x_{n1}, \ldots, x_{ni}), \quad \text{where,} \quad x_n \in \mathbb{R}^{d_1}\)

Aggregation:
\(s(x_k, \theta_\alpha) = \sum_{i} \theta_{\alpha i} \phi_i(x_k), \quad \text{where} \quad \theta_{\alpha i} \in \mathbb{R}^{d_1}\)

Part-based representation:
\(f(x, \theta, W) = [f(x, \theta_1, W), \ldots, f(x, \theta_C, W)]\)

Final Classifiers:
\(\text{if} \quad w_j > \text{threshold} \quad \text{then} \quad \text{Classify} \)

Score of \(\alpha\)-th part classifier for region \(x_k\):
\(s(x_k, \theta_\alpha) = \sum_{i} \theta_{\alpha i} \phi_i(x_k), \quad \text{where} \quad \theta_{\alpha i} \in \mathbb{R}^{d_1}\)

Let:
\(\Theta = [\theta_1, \ldots, \theta_C], \quad p = 1 \ldots P\)

Multi-Class Classification Loss:
Using Logistic regression, class label for an input image \(X\) is predicted as:
\[ P(Y=y|x, \Theta, W, \beta) = \frac{\exp \left( \sum_{i} \theta_{yi} \phi_i(x) \right)}{1 + \sum_{j} \exp \left( \sum_{i} \theta_{ji} \phi_i(x) \right)} \]

Motivation

Motivation for using PBM:
1. By breaking object/image into parts – all of them are visible and recognizable.
2. Parts can be recomposed in many ways to construct an object/scene.
3. Parts are distinctive of a particular class.
4. Part-based representation is compact.

Challenges:
1. How to train the part classifiers?
2. Difficult problem – No annotations of relevant regions within an image.
3. Chicken and egg problem: Discriminative vs. Selective search.
4. Initialization critical.

Approach

Novel joint learning framework for learning parameters in PBM:
1. We describe each part classifier as a linear combination of weak non-linear classifiers.
2. We introduce a parameter, referred as the “per-part softness pooling coefficient” inside the optimization process.
3. Learning is done by a block-wise Stochastic Gradient Descent (SGD).

Contributions:
1. A boosted non-linear part classifiers.
2. Parametric soft-max Pooling.

Normalised weights:
\[ \nu_j^\alpha = \exp \left( \beta j \phi_j(x_k) \right) \sum_{i} \nu_i^\alpha = 1 \]

Score of \(\alpha\)-th part classifier for region \(x_k\):
\[ s(x_k, \theta_\alpha) = \sum_{i} \theta_{\alpha i} \phi_i(x_k), \quad \text{where} \quad \theta_{\alpha i} \in \mathbb{R}^{d_1}\]

Let:
\(\Theta = [\theta_1, \ldots, \theta_C], \quad p = 1 \ldots P\)

Experimental Settings:

Datasets:
1. Pascal-VOC-2007 (20 classes) - animals (e.g., dog, cat), vehicles (e.g., aeroplane, car) and other manufactured objects (e.g., tv monitor, chair).
2. MIT Indoor67 (67 classes) - e.g., nursery, movie theater, casino or meeting room.
3. Willow dataset (7 classes) - e.g., play instrument, walk.

Performance measure:
1. mean Average Precision (mAP) for Pascal-VOC and Willow.
2. Accuracy (acc) for MIT dataset.

Region proposal scheme:
Selective search.

Region feature extraction:
1. 128-D from 13-layer architecture (VGG-128)
2. 4096-D from 16-layer architecture (VD-16)

Comparison with state-of-the-art methods:

Pascal-VOC-2007

Mit In-door67 (left) and Willow (right)

Methods

VGG128-G

VGG128-G

Oquab et al.

75.5

81.72

Li et al.

85.12

Cimpoi et al.

85.81

Gong et al.

88.47

Region

VD-16-G

85.58

VD-16-G

88.01

Pascal-VOC-2007

MIT Indoor67 (left) and Willow (right)

VGG128-G

84.68

VGG128-G

85.67

Region

VD-16-G

85.12

VD-16-G

88.47

SPLeaP-VD-16-G

88.01

SPLeaP-VD-16-G

85.67

13-layer architecture

1. Parametric Soft-Max Pooling:

Average pooling
Max Pooling
Cross validation
Learned \(j_p\)

80.77

83.23

84.31

84.68

2. A boosted non-linear part classifiers:

Specialization of Part Classifiers to discriminative regions:

Heatmaps for classes:
1. “potted plant”, “bird”, “bottle” and “TV monitor”.

Selectivity of Part Classifiers using heatmaps:

Discriminative parts for the four classes:
1. “horse”, “motorbike”, “dining table”, and “potted plant”.

Conclusion

1. We introduce SPLeaP, a novel part-based model for image classification.
2. Based on non-linear part classifiers combined with part-dependent soft pooling - both being trained jointly with the image classifiers surpasses standard pooling approaches and other PBM on several challenging classification tasks.
3. Our method does not need any particular initialisation of the parts.

References

Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation

Qi Ye*, Shanxin Yuan*, Tae-Kyun Kim
{q.ye14, s.yuan14, tk.kim}@imperial.ac.uk

Motivation

- Existing hierarchical methods mainly focus on the decomposition of the output space while the input space remains almost the same along the hierarchy.
- The spatial attention mechanism is proposed to integrate cascaded and hierarchical regression into a CNN framework by transforming both the input and feature space and the output space, which greatly reduces the viewpoint and articulation variations.
- Between the levels in the hierarchy, the hierarchical PSO forces the kinematic constraints to the results of the CNNs.

Structure

- The Spatial Attention Mechanism integrates the cascaded and hierarchical hand pose estimation into one framework.
- The hand pose is estimated layer by layer in the order of the articulation complexity, with the spatial attention module to transform the input, feature and output space.
- Within each layer, the partial pose is iteratively refined both in viewpoint and location with the spatial attention module, which leads both the feature and output space to a canonical one.
- After the refinement, the partial PSO is applied to select estimations within the hand kinematic constraints among the results of the cascaded estimation.

Kinematic Constraint

- Energy function for each layer
  \[ E_p(\theta^i) = P(\varphi^i)Q(\varphi^i) \]
  - prior probability of the \( i \) sample belonging to the Gaussian distribution centered on the estimation results from CNN
  \[ Q(\varphi^i) \propto \sum_s e^{-\frac{1}{2} k(\varphi^i - \mu_s)} \]
  - force each joint to lie inside the hand silhouette and inside the depth range of a major point cloud.

Results

- Errors for a joint of 4 cascaded stages; In-plane viewpoint distribution of testing set for different stages on ICVL, NYU and MSRC datasets
- Self comparison: demonstrate the impact of different strategies by topping them up to the baselines: Cascade, Spatial Attention, Hierarchy, Kinematic Constraint
- Comparison with state-of-the-art methods
VolumeDeform:
Real-time Volumetric Non-rigid Reconstruction
Matthias Innmann¹, Michael Zollhöfer², Matthias Nießner³, Christian Theobalt², Marc Stamminger¹
¹University of Erlangen-Nuremberg, ²Max-Planck-Institute for Informatics, ³Stanford University

ABSTRACT

We present a novel approach for the reconstruction of dynamic geometric shapes using a single hand-held consumer-grade RGB-D sensor at real-time rates. Our method builds up the scene model from scratch during the scanning process, thus it does not require a pre-defined shape template to start with. Geometry and motion are parameterized in a unified manner by a volumetric representation that encodes a distance field of the surface geometry as well as the non-rigid space deformation. The problem is tackled in real-time at the camera’s capture rate using a data-parallel flip-flop optimization strategy. Our results demonstrate robust tracking even for fast motion and scenes that lack geometric features.

METHOD OVERVIEW

First, a deformed 3D mesh is extracted from the signed distance field using Marching Cubes. The mesh is rendered to obtain a depth map, which is used to generate dense depth correspondences. Next, we match SIFT features of the current frame with those of all previous frames. Based on all correspondences, we optimize the deformation field such that the resulting model explains the current depth and color observation. Finally, we integrate the RGB-D data of the current frame.

DEFORMATION ENERGY

To update the deformation field, two distinct and complementary types of correspondences between the current deformed shape and the new color and depth input are searched: for depth-image alignment, we perform a fast data-parallel projective lookup to obtain dense depth correspondences. Since in many situations depth features are not sufficient for robust tracking, we also use color information, and extract a sparse set of robust color feature correspondences. These also serve as global anchor points, since their descriptors are not modified over time. To reconstruct non-rigid surfaces in real time, we have to update the space deformation at sensor rate.

For simplicity of notation, we stack all unknowns of local deformations in a single vector:

\[
X = [\ldots, \mathbf{t}_i^T, \ldots, \mathbf{R}_i^T, \ldots]^T.
\]

To achieve real-time performance, even for high-resolution grids, we cast finding the best parameters as a non-linear variational optimization problem. Based on these definitions, we define the following highly non-linear registration objective:

\[
E_{\text{total}}(X) = w_\text{data} E_{\text{data}}(X) + w_\text{prior} E_{\text{prior}}(X),
\]

where

\[
E_{\text{data}}(X) = \sum_{i=1}^{N} w_i \left( \|S(\mathbf{p}_i) - \mathbf{p}_i^T\|^2 \right)^{\frac{1}{2}},
\]

\[
E_{\text{prior}}(X) = \sum_{i=1}^{N} \|\mathbf{R}_i \mathbf{t}_i - \mathbf{t}_i\|^2.
\]

PARALLEL ENERGY OPTIMIZATION

The non-linear optimization objective \(E_{\text{total}}\) can be split into two independent subproblems by employing an iterative flip-flop optimization strategy: first, the rotations \(\mathbf{R}_i\) are fixed and we optimize for the best positions \(\mathbf{t}_i\). Second, the positions \(\mathbf{t}_i\) are considered constant and the rotations \(\mathbf{R}_i\) are updated. These two step are iterated until convergence.

1. We find the optimal positions by solving the linear system \((L + B^T B) \mathbf{t} = b\).

2. We obtain the best fitting rotation based on Procrustes analysis with respect to the canonical pose. With our implementation, we can compute the best rotations for 400K voxels in 1.9ms.

RESULTS

The following figure demonstrates the importance of our sparse color tracker:

To update the deformed 3D mesh, we extract SIFT features of the current frame with those of all previous frames. Based on all correspondences, we optimize the deformation field such that the resulting model explains the current depth and color observation. Finally, we integrate the RGB-D data of the current frame.

COMPARISON TO STATE-OF-THE-ART

We compare our method to a state-of-the-art template based approach. The results are comparable – without the necessity of a pre-scanned template.
Introduction

• Existing monocular vSLAM methods face difficulties that include:
  - presence of outliers
  - dynamic foregrounds and pure rotation of the camera
  - large baselines
  - scale drift
  - density of 3D reconstruction
  - computational efficiency

• Using planes instead of point features is advantageous in SfM frameworks.

• A recent approach [1] has shown how to quickly segment affine correspondences (ACs) that belong to the same plane.

• We propose nMatch, a complete vSLAM pipeline that relies on plane features to estimate the camera motion and provide a PPR of the scene.

Experiments

2-View Experiments

• 3 pairs of images from the KITTI dataset:
  - Example 1: normal camera motion
  - Example 2: dominant dynamic foreground
  - Example 3: static camera

• Scale estimation: Points are reconstructed for every two pairs of frames and the scale is the median of the per-point ratios

• Discrete optimization: sliding window approach for back-propagating planes across frames, improving the accuracy and visual perception of the 3D model

Large-Scale Experiments

• 4 sequences of the KITTI dataset:
  - Seq. 1: 125 frames
  - Seq. 2: 268 frames
  - Seq. 3: 395 frames
  - Seq. 4: 1100 frames

References:
**Peripheral Expansion of Depth Information via Layout Estimation With Fisheye Camera**

Alejandro Perez-Yus, Gonzalo Lopez-Nicolas, Jose J. Guerrero

[alperez, gonlopez, josechu.guerrero]@unizar.es

Instituto de Investigación en Ingeniería de Aragón (I3A) - Universidad de Zaragoza

**Motivation**

Most consumer RGB-D cameras have a field of view (FoV) too small for certain applications.

On the other hand, there are many cameras (such as fisheye cameras) which are able to capture color images with a large FoV, but lacking the 3D information.

**Proposal**

New hybrid system with fisheye and depth cameras to overcome the limitations, having:

- Depth certainty and scale
- Wide field of view (>180 deg)

To use such system we propose to extend the depth information to the fisheye image via layout estimation. At the end we get a new 180° depth image where the center has the initial depth information and the periphery a good estimation of the structure of the room.

**Method**

**A. Calibration**

The system requires to be calibrated to map the depth information to the fisheye camera frame. Extended information of the procedure in [1].

**B. Planes, Lines and Vanishing Points**

Planes are extracted from the depth information finding large areas of points with normals in the three dominant directions (initial computation of the Vanishing Points, VP). Floor plane is extracted to provide scale.

Lines are extracted using [2] and used to refine the position of the VPs. Line segments are classified and associated to depth intersections.

C. Floor Plan Projection

Horizontal line segments under horizon are projected to the floor plane to form a scaled contour distribution similar to a floor plan. This type of projection is used to find the height of the ceiling assuming floor-ceiling symmetry.

**D. Corner Extraction**

We define four types of corner intersections with only two lines. Floor and ceiling lines are equally considered.

Corners are scored to make more relevant for further stages those with more lines, longer, closer to the intersection and those associated to 3D lines.

**E. Hypotheses Generation**

From the set of corners we select a random number of them increasing the probability to be picked by the score.

We generate physically-coherent closed Manhattan layouts with a similar procedure to the example shown in the figure.

Since we only consider Manhattan-valid layouts, it is possible to recover undetected corners when two consecutive corners are not following Manhattan convention.

**F. Hypotheses Evaluation**

For each hypotheses (a) we can generate floor plans (b), virtual images showing floor/wall distribution (c), depth images (d) or point clouds (e-f).

For the evaluation, we propose three methods:

- Sum of Scores (SS) of the corners involved in the hypotheses.
- Sum of Edges (SE) of the layout overlapping in the floor plan contours.
- Orientation Map (OM) [3].

The quality of the solutions are measured with the Mean Accuracy (%) showing the right per-pixel wall/floor distribution overlap between layout hypotheses and ground truth. From our own dataset of 70 captures we go the following numerical results:

**References**


**Acknowledgments**

This work was supported by Projects DPI2014-61792-EXP and DPI2015-65962-R (MINECO/FEDER, UE) and grant BES-2013-065834 (MINECO).
Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation

Fatemehsadat Saleh1,2, M. Sadegh Ali Akbarian1,2, Mathieu Salzmann3, Lars Petersson1,2, Stephen Gould1, and Jose M. Alvarez1,2
1The Australian National University (ANU) 2Data61-CSIRO, Australia 3CVLab, EPFL, Switzerland

Introduction

Goal: Assigning a semantic label to every pixel in the image.

Problem: Acquiring a huge amount of pixel-level annotations is expensive. Our approach: We aim at using one of the weakest levels of annotation, image-level tags. Drawbacks of current approaches using image-level tags:
• Poor localization and inaccurate object boundaries.
• Additional priors require pixel-level annotations/bounding boxes. Different types of annotation used in related works

Contributions
• A method to extract accurate masks from a network pre-trained for object recognition.
• A novel loss function to incorporate these masks during training.
• A method to extract accurate masks from a network pre-trained for object recognition.

Our Method

A new form of weak supervision, where the user selects the best mask among several automatically generated candidates.

Built-in Foreground/Background Model
From a network pre-trained on ImageNet, we propose to exploit the unit activations of the hidden layers to extract a foreground/background mask.

Benefits
Foreground/background mask extracted without relying on an external method. Requires no additional annotations.

Novel Loss Function

\[
S^f_i = \frac{1}{r} \log \left[ \frac{1}{|M|} \sum_{j \in M} e^{S^f_j} \right] \\
L_{\text{mask}} = -\frac{1}{|L|-1} \sum_{k \in L, k \neq 0} \log(S^f) - \log(S^f) - \frac{1}{|L|} \sum_{k \in L, k \neq 0} \log(1 - S^f_k)
\]

References:

For further information
Fatemehsadat Saleh
E: fatemehsadat.saleh@data61.csiro.au
Go to Goal

segmenting a video into static environment and moving objects.
- modeling camera motion and object motion
- Bayesian approach: new flow angle likelihood for assigning pixels to a motion model

Camera Motion Field

When the camera moves, pixels belonging to the static background no longer maintain their position in consecutive frames. They move accordingly to a translational or/and rotational motion field due to camera motion.

Angle Field

Top to bottom: the original frame, the observed translational angle field, the best fitting translational camera angle field and the segmentation results using our motion segmentation method. The moving object is shown in red.

Modeling Motion

Left to right: Frame 1 of the video sequence cars2 of the BMS-26 data set; binary ground truth showing static environment in black and moving objects in white and observed translational angle field described by four different motion models.

Three of the motion models describe the three differently moving objects the fourth motion model describes the pixel displacement of the static environment due to camera translation.

Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Camfragge MCC</td>
<td>0.4305</td>
<td>0.3517</td>
<td>0.1633</td>
<td>0.3354</td>
<td>-</td>
<td>0.5344</td>
</tr>
<tr>
<td>F</td>
<td>0.4379</td>
<td>0.3297</td>
<td>0.1602</td>
<td>0.3007</td>
<td>-</td>
<td>0.5276</td>
</tr>
<tr>
<td>BMS-26 MCC</td>
<td>0.6851</td>
<td>0.6112</td>
<td>0.7187</td>
<td>0.6399</td>
<td>-</td>
<td>0.7576</td>
</tr>
<tr>
<td>F</td>
<td>0.7306</td>
<td>0.6412</td>
<td>0.7276</td>
<td>0.6595</td>
<td>0.6246</td>
<td>0.7823</td>
</tr>
<tr>
<td>ComplexBq MCC</td>
<td>0.4752</td>
<td>0.6359</td>
<td>0.3257</td>
<td>0.3661</td>
<td>-</td>
<td>0.7491</td>
</tr>
<tr>
<td>F</td>
<td>0.4559</td>
<td>0.6220</td>
<td>0.3300</td>
<td>0.3297</td>
<td>0.3751</td>
<td>0.7408</td>
</tr>
<tr>
<td>Total avg. MCC</td>
<td>0.5737</td>
<td>0.5375</td>
<td>0.4866</td>
<td>0.5029</td>
<td>-</td>
<td>0.6918</td>
</tr>
<tr>
<td>F</td>
<td>0.5970</td>
<td>0.5446</td>
<td>0.4911</td>
<td>0.4969</td>
<td>-</td>
<td>0.6990</td>
</tr>
</tbody>
</table>

Comparison to state-of-the-art: Matthew’s correlation coefficient and F-measure for each method and data set.

Bayesian Model

Bayes’ rule
Posterior probabilities of each motion model $M_j$ at each pixel location

$$p(M_j | v_k) \propto p(v_k | M_j) \cdot p(M_j)$$

Flow Angle Likelihood

The von Mises distribution. When a motion field vector $v_k$ is perturbed by added Gaussian noise, the distribution over optical flow angles is well-modeled by a von Mises distribution. The figure shows the best von Mises fit to these sample distributions and the blue curve shows the lower quality of the best Gaussian fit.

Prior
We create a prior at each pixel for each motion model in the new frame by propagating the posterior from the previous frame.

$$p(\theta_{v_k} | v_k, M_j) \propto \mathcal{N}(\theta_{v_k} | \mu = \theta_{v_k} \cdot \kappa = \alpha \| v_k \|)$$

$\alpha$ and $\beta$ are parameters, that add flexibility to that model.

Method

We model the flow likelihoods using a von Mises distribution with parameters $\mu$ and $\kappa$, where $\kappa$ depends on the flow magnitude $\|v_k\|$.

$$p(\theta_{v_k} | v_k, M_j) \propto \mathcal{M}(\theta_{v_k} | \mu = \theta_{v_k} \cdot \kappa = \alpha \| v_k \|)$$

Magnitudes are specifying the informativeness of the flow angles.

Posteriors are propagated from the previous frame to the new frame.

Model: Motion direction of a motion component projected on the image plane angle field.

Vector $v_k$ containing only motion due to camera translation and object motion.

$M$: flow vector describing the motion field caused by a translating camera in its stationary 3D environment. $M$ does not include motion of independently moving objects.

The translational magnitudes alone have no information about which motion is most likely.
**Motivation**

Early works:
- Pixel-level: exploit intensity, color, orientation, texture, etc
- Region-level: integrate region and contextual information

Problem:
- Both should consider the relationship between image elements from overall and local perspectives
- Ignore the semantic completeness of salient objects

Solution:
- Utilize category-independent object characteristics of region proposals
- View saliency detection as a ranking problem

**Joint Ranking and Subspace Learning**

Joint learning formulation:
\[
E = \frac{1}{2} ||w||^2 + \lambda \sum_{(i,j) \in P} \max(0, 1 - w^T P(k_i - k_j))^2 + \mu \sum_{n=1}^K I_n(1 - \frac{||P(k_n) - k_n||^2 - 1)} + \mu Tr(PKP^T)
\]

Iterative optimization:
(A) Fix P, update the ranking weight w using Truncated Newton optimization
(B) Fix w, update the projection matrix P using gradient descent

Final Saliency Map:
\[
S(x) = \sum_{i=1}^K \exp(2 \times s_i) \times m_i(x)
\]

where \(s_i = w^T P k_i\) is the ranking score of \(i\)-th proposal

**Ranking Learning**
Encourage foreground proposals should have higher ranking scores than background proposals

**Subspace Learning**
Narrow the distances between similar samples and simultaneously expand the gaps of dissimilar samples

**Experiments**

- **P-R curve**
  - PASCAL-S dataset
  - MSRA dataset
- **F-measure**
  - SOD dataset
  - ECSSD dataset
- **AUC**
  - Dataset: PASCAL-S, MSRA, SOD, ECSSD

**Visual comparison**

**Acknowledge**

The work was supported by the National Natural Science Foundation of China under Grant #61371157, Grant #61472060 and Grant #61528101.
Leaking” into background affects measurement too

Application on and RGB+D dataset for tremors measurement

Use depth cue to magnify motion

Per frame, build complex steerable pyramid

Generalize the Fast Bilateral Filter

Depth-Aware Steerable Pyramids

Compare to accelerometer on chest and hand

Standard pyramid: measures leaked motion of hand

Magnification comparison to state-of-the-art

Our approach

Using depth-aware pyramid representation curbs artifacts

More single frame magnification comparisons

Contributions

- Use depth cue to magnify motion of occluded regions
- Depth-Aware Steerable Pyramids
- Generalize the Fast Bilateral Filter to Non-Gaussian bilateral filters
- Application on and RGB+D dataset for tremor measurements

Main motivation

Medical application of full body tremor assessment
- Real-world hospital setting (e.g. Parkinson patients)
- Need to discover and measure small motions in arms, body, head, with minimum patient effort
- Should be robust against viewpoint, self-occlusions, and presence of large motions
Other uses of our novel filter explored in Sup. Mat.

Example magnification task

Magnify small motions in body, but large movements in foreground

Building a steerable pyramid

Our novel non-Gaussian bilateral filter generalizes the Fast Bilateral Filter [3]

- Given input image I(x), depth image E(x), and let x, y, z be 2D image locations
- The standard bilateral filter outputs O(x), using Gaussian kernel G(d; σ)
  \[ O(x) = \frac{1}{W(x)} \sum_{y \in N(x)} W(|x - y|, E(x) - E(y)) I(y) \]
  \[ w(d_x, d_y) = G(d_x; \sigma_x) \times G(d_y; \sigma_y) \]
- Our non-Gaussian bilateral filter Q(x) for non-Gaussian kernels F(d)
  \[ Q(x) = \sum_{y \in N(x)} F(|x - y|) O^+(x, E(x)) \]
  Here \( O^+(x, \xi) \) is a volumetric representation (2D image + 1D depth)
  \( Q \) convolves \( F \) on all depth layers of \( O^+ \), which locally downweights the input at distant depth layers through a 3D kernel \( w \)

Processing pipeline comparison

Steerable Pyramids also used for motion measurement

- “Leaking” into background affects measurement too
- Using our bilateral pyramid is therefore more robust

Task: measure the chest area behind the moving hand
- Compare to accelerometer on chest and hand
- Standard pyramid: measures leaked motion of hand
- Bilateral pyramid: measures chest like accelerometer

Measuring motion task

Non-Gaussian bilateral experiments

Study non-Gaussian filters on images + binary mask
- Ideally, filter ignores intensity within masked region
- Compare our method to using inpainting techniques
- Tested on steerable filters and ConvNet filters

References, acknowledgements, and code


Acknowledgments: This work is part of the research programme Technology in Motion (TUM [628.004.001]), financed by the Netherlands Organisation for Scientific Research (NWO)

Code: github.com/jkooij/depthaware-momag
Project page: tim.lumc.nl
Stacked Hourglass Networks for Human Pose Estimation
Alejandro Newell, Kaiyu Yang, and Jia Deng
Computer Science and Engineering, University of Michigan, Ann Arbor

Human Pose Estimation
- From a single RGB image, identify the pixel location of key joints of the body.
- Useful in a variety of applications and critical to expanding our understanding of people in images and video.
- We introduce a novel convolutional neural network (CNN) to address this task outperforming current state-of-the-art methods [1][2].

Related Work
- CNNs + Graphical models: Often used to enforce anatomic constraints between joints. (e.g. Chen et al. NIPS'14, Tompson et al. CVPR'15)
- Methods to refine predictions: Iterative prediction (Carreira et al. CVPR'16), cascades (Tompson et al. CVPR'15), and intermediate supervision (Wei et al. CVPR'16)
- Generating pixel-wise output: Fully convolutional networks (Long et al. 2015), conv-deconv architectures, and many others. These methods may omit top-down processing or require special deconv/uppooling operations.

Hourglass Module
Key ideas:
- Consolidate global and local features
- Even emphasis placed on top-down and bottom-up processing

Implementation details:
- Each box below represents a residual module (He et al) composed of three convolutional layers (1x1-128, 3x3-128, 1x1-256)
- Pixel-wise predictions are produced by applying two rounds of 1x1 convolutions to the final set of features (1x1-256, 1x1- # of joints)
- Simple process to combine features across scales: nearest-neighbor upsampling and elementwise addition [1]
- Skip connections retain high resolution information lost by pooling

Stacked Hourglass Architecture
- Hourglasses are placed consecutively; the features of one fed as input into the next. Weights are not shared between hourglasses.
- Stacking allows for repeated bottom-up, top-down inference
- Predictions are generated after each hourglass, and a loss applied.

- These predictions are reintroduced back into the network (visualized above) allowing the network to reassess estimates and shift predictions towards a more coherent global state.
- Residual links are included across hourglass modules (dotted line)

Experiments
- We compare a stacked network to a single extended hourglass with the same number of layers and approximately same number of parameters. (top right)
- Stacking hourglasses with intermediate supervision improves training performance and rate of convergence.
- Intermediate supervision at lower resolutions for the single hourglass model does not offer the same benefits.
- Taking stacking even further, we compare a 2, 4, and 8-stack network showing that rearranging for more hourglasses further improves performance. (bottom right)

Comparisons of Intermediate/Finial predictions. Child vs 8th hourglass output on validation image

Results
Benchmark datasets:
Frames Labeled in Cinema (FLIC) [3] - 5k images
MPII Human Pose [4] - 25k images, ~40k annotated people

References

Full paper and code available at: www-personal.umich.edu/~shnewell/pose
Real-time Large-Scale Dense 3D Reconstruction with Loop Closure

Olaf Kähler, Victor A Prisacariu, David W Murray
University of Oxford, Active Vision Lab
http://www.infinitam.org

Task

Aim: track and reconstruct large spaces densely
Problem: tracking drifts – errors on loop closure
- Recognise previously visited locations
- Globally adjust previous estimations
For online dense SLAM this is impractical
- Adjusting large scenes becomes slow
- Existing solutions require offline post-processing

Scene represented by collection of submaps
Use volumetric fusion for each submap independently
Relative constraints between submaps gathered online
On loop closure, adjust graph of submaps
→ Overall processing takes milliseconds per frame

Our Solution

System Overview

Core Engine - InfiniTAM

Find constraints between submaps:
- Track same image in multiple submaps i and j: Poses \( T_{ij} \)
- Pose between submaps: \( T_{ij} = T_{ij}^{-1} T_u \)
- Robustly aggregate over time \( t \) to get final estimate \( T_u \)

Key idea: Maintain list of several active submaps:
- Start with standard fusion pipeline for a single submap
- Once initial part is out of view, start new submap
- Track both submaps simultaneously and accumulate relative constraints
- Once relocaliser detects loop closure, start new standard tracking pipeline
- Run graph optimisation with accumulated constraints
- Visualisation: On-the-fly fusion of all submaps

Processing time:
- 7.1 – 8.5 ms per frame
- Remains constant

Pose Graph Optimisation

Find global pose for each submap such that relative constraints are satisfied
For visualisation, submaps are fused on-the-fly with best estimate of global pose
Pixel-Level Domain Transfer

**Donggeun Yoo**, Nam Il Kim, Sunggyun Park, Anthony S. Paek, In So Kweon. KAIST, Lunit Inc.

**Motivation**
- Put mental imagery into AI.
- Train a converter \( C(I_S|\Theta^C) \).

\[ \hat{I}_T = C(I_S|\Theta^C) \]

**Contributions**
- Proposing the first framework for semantically transferring a source knowledge to a target domain in pixel-level.
- Proposing a novel discriminator that enables us to train the semantic relation between domains.
- Building a large clothing dataset expected to contribute to domain adaptation researches.

**Architecture**

- **Source domain**
  - Encoder
  - \( I_S \)
  - \( 64 \times 64 \)
  - \( 1 \rightarrow 32 \rightarrow 64 \)
  - \( 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \) \( \rightarrow 1024 \) \( \rightarrow 1024 \)
  - \( 1024 \) \( \rightarrow 512 \) \( \rightarrow 256 \) \( \rightarrow 64 \) \( \rightarrow 64 \)

- **Converter**
  - \( C \)
  - \( 3 \rightarrow 64 \)

- **Target domain**
  - Decoder
  - \( I_T \)
  - \( 64 \times 64 \)
  - \( 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \) \( \rightarrow 1024 \) \( \rightarrow 1024 \)
  - \( 1024 \) \( \rightarrow 256 \) \( \rightarrow 128 \) \( \rightarrow 64 \) \( \rightarrow 64 \)

- **Real/fake-discriminator** \( D_R \)
  - \( I_T \) or \( \hat{I}_T \) or \( \hat{I}_T \)

- **Domain-discriminator** \( D_A \)
  - \( I_S \)
  - \( 64 \times 64 \)

**LookBook Dataset**
- **Fully paired** 77,546 images in total.
  - 8,726 product images.
  - 68,820 fashion model images.
  (Available at [https://dgyoo.github.io/](https://dgyoo.github.io/))

**Generation Results**

**Quantitative evaluation**

<table>
<thead>
<tr>
<th>Methods</th>
<th>User study score</th>
<th>Pixel-level (dis)similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Realistic</td>
<td>Attribute</td>
</tr>
<tr>
<td>C+RF</td>
<td>0.40</td>
<td>0.21</td>
</tr>
<tr>
<td>C+MSE</td>
<td>0.28</td>
<td>0.60</td>
</tr>
<tr>
<td>C+RF+DD (Ours)</td>
<td><strong>0.82</strong></td>
<td><strong>0.67</strong></td>
</tr>
</tbody>
</table>

**Less-variant to varying source conditions**

**Human → product generation**

**Product → human generation**
Abstract

Aiming at accelerating the test time of deep convolutional neural networks (CNNs), we propose a model compression method that contains a novel dominant kernel (DK) and a new training method called knowledge pre-regression (KP). In the combined model DK²PNet, DK is presented to significantly accomplish a low-rank decomposition of convolutional kernels, while KP is employed to transfer knowledge of intermediate hidden layers from a larger teacher network to its compressed student network on the basis of a cross entropy loss function instead of previous Euclidean distance. Compared to the latest results, the experimental results achieved on CIFAR-10, CIFAR-100, MNIST, and SVHN benchmarks show that our DK²PNet method has the best performance in the light of being close to the state of the art accuracy and requiring dramatically fewer number of model parameters.

Introduction

In recent years, deep convolutional neural network (CNN) has made impressive success in several computer vision tasks such as image classification [1], object detection and localization [2,3]. On many benchmark challenges [1,4-6], records have been being consecutively broken with CNNs since 2012 [1] on. Surprising performance, however, usually comes with a heavy computational burden due to the use of deeper and/or wider architectures. Complicated models with numerous parameters may lead to an unacceptable test or inference time consuming for a variety of real applications. To resolve such challenging problem, there was an early interest in hardware-specific optimization [1,7-19]. But it is very likely to be unable to meet increasingly demanding for real-time applications in an era of mobile Internet. The reason is that huge amounts of portable devices such as smart phones and tablets are often equipped with low-end GPUs and CPUs as well as limited memory.

To speed up cumbersome CNNs, one is to directly compress existing large models or ensembles into small and fast-to-execute models [8-11,15]. Another is to employ deep and wide top-performing teacher networks to train shallow and/or thin student networks [16-19]. Based on low rank expansions, convolutional operator is decomposed into two procedures: feature extraction and feature combination (Sect. 3.1). Initially inspired from low rank approximation of responses proposed by Zhang et al. [15], this paper proposes a novel dominant convolution kernel (DK) for greatly compressing filter banks. To deal with performance degradation problems caused by model compression, we present a new knowledge pre-regression (KP) training method for compressed CNN architectures, which expands the FitNet training method [10] to make it much easier to converge and implement. Such KP-based training method fills the intermediate representation gap between original teacher network and compressed student.

Fig. 1. Dominant Convolutional Kernel (DK). (a) Regular convolutional kernel; (b) DK of $1 \times 1$; (c) DK of $1 \times 2$; (d) DK of $1 \times n$.

As for our dominant convolutional kernels shown in Figs. 1(b), 3(d), feature is extracted only for single incoming feature map, while feature combination is done across temporarily convolved maps through $1 \times 1$ weights. Specifically, instead of convolving input feature maps with $c_{i} \times c_{o}$ convolutional kernels in regular convolutional layer, we merely employ $c_{i}$ convolutional kernels, called dominant kernel (DK) in this paper. Apparently, there is 1-to-1 correspondence between inputs and temporarily convolved feature maps. This can be extended to the 1-to-n case, where $1 \leq n \leq k_{d} \times k_{o}$, as shown in Fig. 1(c) and (d). As a result, we have a total of parameters of $n \times n \times k_{d} \times k_{o}$ in such feature extraction phase. During feature combination, each output channel is a weighted sum of all the temporarily convolved maps. Note that weight vectors of $n \times c_{i}$ are exploited for combination of different maps. Thus there are totally $n \times n \times c_{o}$ parameters in such combination procedure. As $n$ reaches up to the maximum of $k_{d} \times k_{o}$, our dominant convolutional layer may have capabilities of approximation similar as regular convolutional layer. It is because linear independence vectors of $k_{d} \times k_{o}$ determines a set of basis vectors for $k_{d} \times k_{o}$-dimensional feature spaces.

Knowledge Pre-regression Training Method

The hint/guided layer pairs, however, usually have different dimensions. Russo et al. [19] deal with this problem by introducing a convolutional regressor to match their dimensions, whose Euclidean loss is utilized to optimize them. Although this method is simple, it is hard to converge. In this manner, we propose a KP-based training method to tackle such problem, as shown in Fig. 2. An auxiliary regressor with a fully-connected layer $R_{g}$ and the ground truth label $l$ is introduced to the hint layer to generate a target probability distribution $h_{i}$. The probability distribution is then transferred to the guided layer. Meanwhile, we add another auxiliary regressor with a fully-connected layer $R_{g}$ for the guided layer of student network. Furthermore, we adopt the cross entropy loss function to connect the hint and the guided layers. The loss function in our KP-based training method can be written as Eqs. (1), (2), and (3),

$$L_{CP} (W_{s}, R_{g}, R_{h}) = \mathcal{N}(h_{i}^{c}, \mu_{i}^{c}) + \mathcal{S}(l, h_{i})$$

$$\mu_{i}^{c} = \text{softmax} \left( h_{i}^{c} W_{t}, R_{g} \right) / \tau$$

$$\mu_{i}^{c} = \text{softmax} \left( g(x_{i}, W_{s}, R_{g}) / \tau \right)$$

where $\mathcal{S}$ and $\mathcal{H}$ represent the loss function of softmax and cross entropy, respectively, $h$ and $g$ are two functions that map an input image $x$ to the hint and the guided layers, respectively, $W_{t}$ and $W_{s}$ stand for parameters in teacher network and student network, respectively, a temperature parameter $\tau$ [19] is utilized to soften the probability distribution over classes, and $\lambda$ is a hyper-parameter used to balance these two loss functions.

Experimental Results

Finally, our experimental results show that the proposed DK²PNet provides near state-of-the-art test errors, while requiring notably fewer parameters than regular CNN models. For example, without any data augmentation, the DK²PNet-160 yields the best performance of 7.60% on CIFAR-10 using almost 3.8 times less parameters, when compared to existing state of the art results of 7.62% [29]. On CIFAR-100, while the DK²PNet-160 achieves better performance with a test error of 31.39% (that is close to [30], it only requires roughly 90.2 times fewer parameters. On MNIST without any data augmentation, the DK²PNet-96 with 0.18 million parameters, which is the least number of parameters adopted, obtains a test error of 0.31%. Our DK²PNet-160 receives near state of the art result of 1.83% on SVHN benchmark with roughly 12% of the parameters in comparison with [29], dramatically reducing the number of parameters by a factor of 8.

Acknowledgements

This work was supported in part by the National Science Foundation of China (NSFC) under Grant Nos. 91420106, 60972033, and 60773040, and by the National High-Tech R&D Program of China under Grant No. 2012AAA041402.
Learning Social Etiquette: Human Trajectory Understanding in Crowded Space

Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, Silvio Savarese
Stanford University

Motivation
- Better understanding how human navigate in crowded space
- Capturing their behavior or social etiquette.

Contributions
- New large-scale dataset that collects videos of various type of agents
- New characterization that describes the “social sensitivity”
- New trajectory forecasting method

Training and Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Frames</th>
<th>Targets</th>
<th>Interactions</th>
<th>Hi</th>
<th>Pol</th>
<th>Back</th>
<th>Cal</th>
<th>Car</th>
<th>Bus</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRBOX</td>
<td>2094</td>
<td>6091</td>
<td>3081</td>
<td>102</td>
<td>203</td>
<td>933</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MISSLE</td>
<td>2494</td>
<td>6591</td>
<td>3099</td>
<td>103</td>
<td>204</td>
<td>942</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BROCOLI</td>
<td>2594</td>
<td>6691</td>
<td>3099</td>
<td>103</td>
<td>204</td>
<td>942</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VARIOUS</td>
<td>2694</td>
<td>6791</td>
<td>3099</td>
<td>103</td>
<td>204</td>
<td>942</td>
<td>103</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>7984</td>
<td>25980</td>
<td>9353</td>
<td>303</td>
<td>599</td>
<td>1585</td>
<td>513</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Type of agents: pedestrian, bicycle, car, skateboard, cart, bus

Our Model

Modeling Social Sensitivity

\[ s_d : \text{the distance to the subject to be avoided.} \]

\[ s_w : \text{the radius of influence of other object.} \]

\[ \beta : \text{control the peakness of the weighting function} \]

Forecasting Multiple classes of target

\[ E_d(V(t); n_i, s_d, s_w, \beta) = \sum_{j \neq i} \alpha_j s_d \exp \left( -\frac{d(V_i, V_j)}{s_d^2} \right), \]

\[ s_i = \exp \left( \frac{-d(V_i, V_j)}{s_w^2} \right) \exp \left( \frac{-d(V_i, V_j)}{s_w^2} \right), \]

\[ d(v_i, v_j) = \frac{|v_i - v_j|^2}{v_i^2} \]

Forecasting results from the Multi class dataset

Multi-target tracking

Modified Multi-target Tracking (MTT) algorithm from Xiang et. al. [39].

- Linear forecasting method
- Single-class social force [1]
- Our multi-class forecasting method

Experiments

Evaluation Methods - Forecasting:
- Mean average displacement on all trajectories
- Mean average displacement on collisions avoidance
- Final position average displacement

Forecasting results from the Pedestrians Only Dataset [1]

Conclusions

- New large-scale dataset publicly available
- Better understanding of a target behaviors
- Better prediction and tracking
Introduction

Problem:
Estimate the 3D pose of a subject from a single image

Previous work:
- CNNs [1,2]
- pose-priors [3]
- non-parametric Bayesian [4]

Challenges:
- Ill-posed problem
- View variability/Pose ambiguity
- 2D joint estimation

Contributions:
- 3D pose estimation of each limb/group separately
  - Reference pose (Vitruvian) definition for the 3D features
  - 3D pose estimation of each limb/group separately
  - 2D joint estimation

Dictionary Learning

1st level:
Appearance features based on PHOG around each joint

2nd level:
Pose features based on Principal Geodesic Analysis

References


Results

Number of clusters generated by DPM for each group of joints

Examples of query images and the recovered 3D pose

Average per joint error between estimated 3D poses and ground truth in mm

Contact
Efficient and Robust Semi-supervised Learning over a Sparse-Regularized Graph

Hang Su, Jun Zhu, Zhaozheng Yin, Yinpeng Dong, and Bo Zhang
Department of Computer Science and Technology, Tsinghua University, China
*Department of Computer Science, Missouri University of Science and Technology, USA

Motivation
- Graph-based Semi-Supervised Learning is impressive
- But has limitations
  - Sensitive to parameters
  - Computationally prohibitive
  - Characteristics evolving

Goal: Develop a novel GSSL based on a batch of informative beacons with sparsity appropriately harnessed

Contribution
- Propose an \( \ell_1 \)-Beacon Graph based semi-supervised algorithm
  - Place a batch of characteristic-specific beacons in the feature domain
  - Represent the original samples with a subset of beacons
  - Predict missing labels with label fusion of the corresponding beacons.
- Our proposed method outperforms the previous algorithms
  - Mitigate the computational bottleneck by weighted averaging the soft labels of a subset of beacons
  - Enhance robustness by incorporating the sparse regularization
  - Boost the performance incrementally by expanding the beacon set and update their characteristic parameters dynamically

Construction of \( \ell_1 \)-Beacon Graph
- Generate a set of beacons automatically, which behave as indicators to guide the inference procedure

\[
(\Psi^*, Z^*) = \arg\min_{\Psi, Z} R_\Psi(\Psi, Z) + R_\Omega(\Psi, Z) + \lambda R_\Omega(\Psi, Z)
\]

\( \Psi \): Feature Domain
\( Z \): Label Domain

Unlabeled set
\( R(\Psi, Z) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i, z_i) 
\)

Labeled set
\( R(\Psi, Z) = \sum_{i=1}^{n} \phi(x_i, z_i) \)

Optimization Algorithm
- Solve by alternately minimizing one variable while keeping the other one fixed

Calculate \( Z \) by fixing \( \Psi \)
\[
Z^* = \arg\min_{Z} \frac{1}{m} \sum_{j=1}^{m} \left[ \frac{1}{k} \sum_{i=1}^{k} \phi(x_{ij}, z_{ij}) + \lambda \sum_{i=1}^{k} \phi_{ij} \right]
\]

Calculate the \( \Psi \) by fixing \( Z \) as
\[
\Psi^* = \arg\min_{\Psi} \left[ \frac{1}{m} \sum_{j=1}^{m} \phi(x_{ij}, z_{ij}) - \frac{1}{k} \sum_{i=1}^{k} \phi_{ij} + \lambda \sum_{i=1}^{k} \phi_{ij} \right]
\]

Inductive Inference
- Implement scalable inference by label fusion

\[
y_i \approx Fz_i \quad \tilde{y}_i = \arg\min_k \| y_i - Bz_i \|
\]

Incremental Update of Beacons in Open-Set
- Training and testing data may exhibit different statistics
- Characteristics of samples may evolve over time
- Expand the beacon set incrementally

\[
\vec{B} \triangleq [B, B_k]
\]

add \( k \) new beacons

Beacon Set Update
- Randomly shuffle \( N \) unfamiliar samples

\[
\min_{\tilde{z}} \frac{1}{N} \sum_{i=1}^{N} \left[ \left( \frac{1}{m} \sum_{j=1}^{m} \phi(x_{ij}, \tilde{z}_{ij}) \right) \right]
\]

Update the beacon set

\[
\vec{B}_k = \vec{B}_{k-1} + 2\delta (\vec{F}_k - \vec{B}_{k-1}, \vec{Z}_k)
\]

\[
\vec{F}_k = \vec{F}_{k-1} + 2\delta (Y - \vec{F}_{k-1}, \vec{Z}_k - \lambda \vec{F}_{k-1} - \vec{Z}_k)
\]

Experimental Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>92.97</td>
<td>92.78</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>72.13</td>
<td>92.21</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>55.13</td>
<td>95.53</td>
</tr>
</tbody>
</table>

\( M < 1 \) is effective in classification and object detection
- Robustness to sub-optimal parameters and reduce time cost
- Handle statistics shift

Conclusion
- Implement semi-supervised inference by coupling the design of informative beacon set and sample-to-beacon relationship
- Orders of magnitude more efficient in computation
- Offer a solution to handle unfamiliar samples or unreliable inference
Introduction
Line features are widely used in many applications, such as 3D reconstruction, image mosaicking, and robot navigation.
Main contribution: More robust and accurate non-coplanar line matching method across views.
New line-points projective invariant (Robustness).
Local homography matching and validation (Accuracy).

Motivations
Existing methods and challenging issues:
1. Use the relative relationship of line positions (Wang’09, López’15).
2. Line descriptors based on textures (Zhang’13).
3. Use epipolar constraints or geometric invariants (Fan’10, Al-Shahri’14).
Key observation: The neighborhood of a line is usually coplanar with the line.

Overview
Use line-points invariants to match local plane areas.
Getting local homography of the areas.
Matching lines by local homography transformations.

Construction of Line-points projective invariant
- Line detection and interest points (SIFT) matching.
- Take intersections of coplanar lines (two lines with very close endpoints) instead of endpoints.
- Determine the gradient directions and neighborhoods (local areas to be matched) of lines.
- Characteristic Number [1]: Projective invariant constructed by points that form a loop.

Matching lines with homography
- Neighborhood-to-neighborhood (Local area) similarity measure:
  1. Take the line and matched interest points to construct invariants.
  2. Measure the similarity of neighborhoods with the median distance of the invariants, makes it robust to wrong point matches.
- Take the points in the matched areas to get homography transformations.
- Take the homography transformations to get candidate line pairs.

Testing images
- ‘One area, one (weighted) vote’ strategy to get final results.

Example Results:
- Low-texture
- Similar textures and similar structures

Compare with state-of-the-art methods
- Compare with [2][3].
- Compare with [4].

The proposed method usually can:
- Get more correct matches. (Robustness)
- Get higher correct rates. (Accuracy)

References

Acknowledgments: Natural Science Foundation of China 61402077, 61033012, 11171052, 61272371, 61003177 and 61328206.
Sparse Representation Based Complete Kernel Marginal Fisher Analysis Framework for Computational Art Painting Categorization

Ajit Puthenputhussery, Qingfeng Liu, Chengjun Liu
New Jersey Institute of Technology

INTRODUCTION

- This paper presents a sparse representation based complete kernel marginal Fisher analysis (SCMFA) framework for categorizing fine art images.
- First, we introduce several Fisher vector based features for feature extraction so as to extract and encode important discriminatory information of the painting image.
- Second, we propose a complete marginal Fisher analysis method so as to extract two kinds of discriminant information, regular and irregular.
- In particular, the regular discriminant features are extracted from the range space of the intraclass compactness using the marginal Fisher discriminant criterion whereas the irregular discriminant features are extracted from the null space of the intraclass compactness using the marginal interclass separability criterion.
- Experimental results on the challenging Painting-91 dataset show that our framework achieves the state-of-the-art performance for fine art painting categorization and outperforms other popular image descriptors and deep learning methods.

FEATURE EXTRACTION USING FUSED FISHER VECTOR FEATURES

![Feature Extraction Diagram]

SPARSE REPRESENTATION BASED COMPLETE KERNEL MARGINAL FISHER ANALYSIS FRAMEWORK

Complete Marginal Fisher Analysis (CMFA) Method:
The motivation for CMFA method is that the traditional MFA method uses a PCA projection in the initial step that may discard the null space of the intraclass compactness which may contain useful discriminatory information.

In the complete kernel marginal Fisher analysis method, the strategy is to split the intraclass compactness into two subspaces namely the range space and null space so as to extract two kinds of discriminant features: regular and irregular discriminant features.

\[
T^r = \arg \max \operatorname{tr}(C^r S^r d^r)
\]

\[
S^r d^r = \lambda S^r C^r
\]

\[
U^r = C^r S^r K
\]

\[
S^r C^r = \lambda S^r C^r
\]

\[
T^r = \arg \max \operatorname{tr}(C^r S^r d^r C^r) = \arg \max \operatorname{tr}(S^r)
\]

\[
U^r = C^r S^r K
\]

Discriminative Sparse Representation Model:
The objective of the discriminative sparse representation model is to integrate the representation criterion with the discriminant criterion in order to enhance the discriminative ability of the proposed method.

\[
\min_{p, s} \sum_{i=1}^{m} \left( ||u_i - D^s ||^2 + \lambda ||s_i ||^2 + \alpha \operatorname{tr}(\delta H_w - (1 - \beta) H_w) \right)
\]

\[
s.t. ||d_i|| \leq 1, (j = 1, 2, ..., r)
\]

EXPERIMENTS

<table>
<thead>
<tr>
<th>Method</th>
<th>Artist Classification</th>
<th>Style Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>42.60</td>
<td>53.20</td>
</tr>
<tr>
<td>LBP</td>
<td>28.50</td>
<td>42.20</td>
</tr>
<tr>
<td>Color-LBP</td>
<td>35.00</td>
<td>47.00</td>
</tr>
<tr>
<td>Color-PHOG</td>
<td>22.80</td>
<td>33.20</td>
</tr>
<tr>
<td>CN-SIFT</td>
<td>44.10</td>
<td>56.70</td>
</tr>
<tr>
<td>RGBSIFT</td>
<td>40.30</td>
<td>47.40</td>
</tr>
<tr>
<td>CLBP</td>
<td>34.70</td>
<td>46.40</td>
</tr>
<tr>
<td>CN</td>
<td>18.10</td>
<td>33.30</td>
</tr>
<tr>
<td>Combined Descriptors</td>
<td>53.10</td>
<td>62.20</td>
</tr>
<tr>
<td>CNN F3</td>
<td>56.40</td>
<td>68.57</td>
</tr>
<tr>
<td>CNN F4</td>
<td>56.33</td>
<td>69.21</td>
</tr>
<tr>
<td>MSCNN-1</td>
<td>58.11</td>
<td>69.67</td>
</tr>
<tr>
<td>MSCNN-2</td>
<td>57.91</td>
<td>70.96</td>
</tr>
<tr>
<td>SCMFA (Proposed)</td>
<td>65.78</td>
<td>73.16</td>
</tr>
</tbody>
</table>

Table 1. Comparison of methods on the Painting-91 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Artist Classification</th>
<th>Style Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFA</td>
<td>59.57</td>
<td>66.79</td>
</tr>
<tr>
<td>SCMFA (Proposed)</td>
<td>65.78</td>
<td>73.16</td>
</tr>
</tbody>
</table>

Table 2. Comparison with marginal Fisher analysis method.

INFLUENCE CLUSTER GRAPHS

Figure 1. Framework of the feature extraction process.

Figure 2. (a) shows the artist influence graph and (b) shows the style influence graph.

Figure 3. The confusion matrix for 13 style categories of the Painting-91 dataset.

REFERENCES

Motivation
- Goal: 3D reconstruction from sparse viewpoint
- Assumptions of previous works:
  - Dense Viewpoints
  - Lambertian and non-uniform albedo
  - Shape Prior
    - Single-view [Kar et al., Aubry et al.]
    - Multi-view [Bao et al.]
    - General lighting condition
  - Dense viewpoints
- Contributions
  - Reconstruction free from assumptions above
  - Unifying single- and multi-view reconstruction
  - 3D convolutional LSTM
  - Attention for viewpoint-specific update

3D-Convolutional LSTM
- Attention for viewpoint-specific update:
  - LSTM cell assigned for each 3D region

3D Convolutional LSTM & GRU
\[ f_t = \sigma(W_t^f x_t + U_t^f h_{t-1} + b_f) \]
\[ i_t = \sigma(W_t^i x_t + U_t^i h_{t-1} + b_i) \]
\[ c_t = f_c \odot c_{t-1} + i_t \odot \tanh(W_t^c x_t + U_t^c h_{t-1} + b_c) \]
\[ h_t = \sigma(c_t \odot i_t) \]

Recurrent Neural Network
- Memory

- Long Short-Term Memory
  \[ f_t = \sigma(W_t^f x_t + U_t^f h_{t-1} + b_f) \]
  \[ i_t = \sigma(W_t^i x_t + U_t^i h_{t-1} + b_i) \]
  \[ c_t = f_c \odot c_{t-1} + i_t \odot \tanh(W_t^c x_t + U_t^c h_{t-1} + b_c) \]
  \[ h_t = \sigma(c_t \odot i_t) \]

- Gated Recurrent Unit
  \[ f_t = \sigma(W_t^f x_t + U_t^f h_{t-1} + b_f) \]
  \[ i_t = \sigma(W_t^i x_t + U_t^i h_{t-1} + b_i) \]
  \[ c_t = f_c \odot c_{t-1} + i_t \odot \tanh(W_t^c x_t + U_t^c h_{t-1} + b_c) \]
  \[ h_t = (1 - h_t) \odot c_{t-1} + h_t \odot \tanh(W_t^c x_t + U_t^c h_t + b_c) \]

Deep Residual GRU/LSTM Network

GRU Input Gate Analysis

Multi-view Evaluations

Multi-view Stereo vs. Ours

References

Acknowledgement
- NSF CAREER grant N-1054127
- Toyota Award #122282
- Korea Foundation for Advanced Studies
- NSF GRFP
### Problem
- This paper addresses real-time incremental Face Tracking

### Motivation
- SDM [1] does not support Incremental Learning (IL)
- Chehra [2] shows that SDM can be trained in parallel, but IL is very slow (~4 fps)

### Contributions
- Continuous Regression revisited (show how to incorporate a data term in CR)
- Cascaded Continuous Regression (show that it performs the same as SDM)
- Incremental Learning for CCR, one order of magnitude faster than SDM
- Fully automatic system state-of-the-art results on 300VW [3]

### Linear Regression
\[
\sum_{j=1}^{M} \sum_{k=1}^{K} \| \mathbf{d}(\mathbf{p}_{j}) - \mathbf{R}(\mathbf{I}_{j}, \mathbf{p}_{j}^{*}) + \delta \mathbf{p}_{j,k} \|^{2}
\]

### Continuous Regression Revisited
- Original continuous regression: i.i.d. uniform sampling distribution – very limited in practice
- We propose to incorporate a data-term which helps us define the sampling volume, in which correlations are allowed:
  \[
  \sum_{j=1}^{M} \int \mathbf{d}(\mathbf{p}_{j}) \delta \mathbf{p} \cdot \mathbf{R}(\mathbf{I}_{j}, \mathbf{p}_{j}^{*} + \delta \mathbf{p}) \|^{2} d\delta \mathbf{p}
  \]

### Cascaded Regression (par-SDM)
\[
\mathbf{R} = \mathbf{YX}^{T} (\mathbf{XX}^{T})^{-1}
\]

### Incremental Learning for par-SDM
- Training set \( \mathbf{R}_{T,S} \)
- Updating set \( \mathbf{R}_{T,U} \)
- \( \mathbf{K} = 10, m = 24, d = 2000 \)

### CCR vs iCCR vs State of the Art
- Fully automatic system
- Category C of 300VW
- The Incremental Learning is crucial to attain State of the Art results!
Problem Statement

Correlation filtering from circulant structure of tracking (Henriques et al. 2015 T-PAMI)

- **Isotropic Response** (the Gaussian shaped response) may fail to reveal the circulant structure due to the discontinuity from the cyclic shifts.
- **Squared Loss** is unable to reliably respond to the drastic appearance changes, e.g., in the presence of occlusions.

Solution: Anisotropy

Basic Idea: Using more robust loss function to learn the correlation filter, resulting in an anisotropic response.

$$\min_{\mathbf{w}} \sum_i \ell(f(\mathbf{x}_i) - y_i) + \lambda \|\mathbf{w}\|_2^2$$

where the loss function $\ell \in \{\ell_1, \ell_1\ell_2, \ell_{2,1}\}$.

Implementation: Introducing relax variables

$$\min_{\mathbf{w}, e} \sum_i \ell(e_i) + \lambda \|\mathbf{w}\|_2^2$$

s.t. $e_i = y_i - f(\mathbf{x}_i)$

and alternately optimizing the variables.

An Example of the Anisotropic Response

Peak values the responses on Basketball

Definition of the Peak Sensitivity:

$$s = \frac{1}{n} \sum_{i=1}^{n} (p_i - \bar{p}_m)^2$$

where $p_i$ and $\bar{p}_m$ denote the peak values of the response in the $i$-th frame and the mean peak values in the $n$ frames, respectively.
**Overview**

Deep Self-Correlation Descriptor for Dense Cross-Modal Correspondence

Seungryong Kim\(^1\), Dongbo Min\(^2\), Stephen Lin\(^3\), and Kwanghoon Sohn\(^1\)

\(^1\) Yonsei University, \(^2\) Chungnam National University, \(^3\) Microsoft Research

---

**Introduction**

1. **Goal**
   - To develop a dense descriptor for matching multi-modal and multi-spectral images
   - To develop an efficient computation model

2. **Objective**
   - Extending SSC descriptor by encoding self-similar structures at multiple levels

**Deep Self-Correlation (DSC) Descriptor**

- **SSC: Single Self-Correlation Descriptor**
  - Computing adaptive self-correlation measure

- **DSC: Deep Self-Correlation Descriptor**
  - Extending SSC descriptor by encoding self-similar structures at multiple levels

**Motivation**

- **Efficient Computation for Dense Description**
  - Cost reformulation and fast edge-aware filtering (EAF)
  - Pre-computation scheme for self-correlation surface

**DSC Description Summary**

**Algorithm 1:** Deep Self-Correlation (DSC) Descriptor

- **Input:** \( f_1, f_2 \), random samples \( x_n \)
- **Output:** DSC descriptor \( DSC(f_1, f_2) \)

**Parameters**
- \( N_{SSC}(N_{SSC}) \): The number of circular pyramid bins (point set)
- \( N_{SSC}(N_{SSC}) \): Points at each level

**Experimental Results and Discussion**

- **Parameter Evaluation**
  - \( LSS \), \( DASC \), \( SSC \), \( DSC \)

- **Cross-modal and Cross-spectral Benchmark**
  - \( RGB \), \( NIR \), different exposure, flash-no-flash, blurring

- **DalI Benchmark**
  - Non-rigid deformation and severe illumination changes

**Conclusion**

- **Deep Self-Correlation (DSC) Descriptor**
  - For establishing dense correspondences between images taken under different imaging modalities

**Contact**

- e-mail: srkim89@yonsei.ac.kr
- homepage: http://diml.yonsei.ac.kr/~srkim/

---

European Conference on Computer Vision (ECCV) 2016
Structured Matching for Phrase Localization

Mingzhe Wang, Mahmoud Azab, Noriyuki Kojima, Rada Mihalcea, Jia Deng
Computer Science and Engineering, University of Michigan, Ann Arbor

Objective loss for a set of bbox-phrase pairs:

VGG FC7 Transformation Matrix
Transformation Matrix

Logistic Regression Parser

Structured Matching

• Denote \( y_{ij} \) as the joint configuration of phrase \( p_i \) and region \( r_j \).
• Denote \( \pi_{ij} \) as the weight of phrase \( p_i \) and region \( r_j \).
• Solve the bipartite matching as a problem of Linear Programming:

Logistic Regression

A woman is sitting down and leaning her head on her hand while another woman is smiling and next to her.

Partial Coreference

• Partial coreference: introduced by possessive pronouns "his", "her" or "its". For example:

Structured Matching

A woman is dressed in Asian garb with a basket of goods on her hip. An instructor is teaching his students how to escape a hold in a self-defense class.

A man is working on his house by repairing the windows.

Experiment Setup

• Dataset: Flickr30K Entities [1].
• 31783 images and 508k regions.
• 500k noun phrases and 70k unique phrases.
• Evaluate with Recall@1 across all phrases.
• A region is true if it overlaps with the ground truth in terms of IoU > 0.5.

Quantitative Results

SUCCESSFUL CASES:

A man and a boy holding microphones.
A man wearing a black helmet riding on a bike.
A woman wearing a black jacket with a woman wearing a black jacket are standing close to each other.

References


Code

https://github.com/mingzhew/structured-matching
Crossing-line Crowd Counting with Two-phase Deep Neural Networks

Zhuoyi Zhao, Hongsheng Li, Rui Zhao, Xiaogang Wang

The overall framework
- A deep Convolutional Neural Network (CNN) is proposed for solving the crossing-line crowd counting problem.
- A two-step training scheme is proposed
  - Phase I: learning per-pixel crowd density and crowd velocity maps
  - Phase II: learning final per-pixel crowd counting map

Pixel-level Supervision Maps
- A pixel-level counting map is defined as the training objective (Phase II objective). Each pixel records how many persons has passed this location along x and y directions at current time.
- Such a counting map could be modeled as element-wise multiplication of a per-pixel crowd density map and crowd velocity map (Phase I objectives).
- The CNN is first trained with the Phase I objectives, and then with the Phase II objective.

New Dataset
We contribute a new dataset for evaluating LOI crowd counting algorithms.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># of scenes</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td># of annotations</td>
<td>5,900</td>
<td>5,900</td>
<td>5,900</td>
<td>5,900</td>
</tr>
<tr>
<td>Data type</td>
<td>video</td>
<td>video</td>
<td>video</td>
<td>temporal slices</td>
</tr>
<tr>
<td>Annotation type</td>
<td>trajectory</td>
<td>trajectory</td>
<td>none</td>
<td>total counts</td>
</tr>
<tr>
<td>Frame resolution</td>
<td>1280×720</td>
<td>1280×720</td>
<td>1280×720</td>
<td>1280×720</td>
</tr>
</tbody>
</table>

*This dataset is not publicly available at the time of publication.

New Dataset

| Scene 1: Street | Scene 2: Alley | Scene 3: Alley | Scene 4: Square | Scene 5: Alley |

<table>
<thead>
<tr>
<th>Scene 1</th>
<th>Scene 2</th>
<th>Scene 3</th>
<th>Scene 4</th>
<th>Scene 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.01/5.59</td>
<td>66.9/83.2</td>
<td>349/404</td>
<td>517/639</td>
<td>511/639</td>
</tr>
<tr>
<td>7.21/14.6</td>
<td>17.9/31.6</td>
<td>398/569</td>
<td>10.3/18.5</td>
<td>7.5/13.9</td>
</tr>
<tr>
<td>13.4/15.1</td>
<td>20.3/33.5</td>
<td>254/362</td>
<td>18.5/14.9</td>
<td>11.8/12.4</td>
</tr>
<tr>
<td>20.7/13.7</td>
<td>17.8/28.0</td>
<td>318/464</td>
<td>24.4/16.7</td>
<td>22.5/19.3</td>
</tr>
<tr>
<td>All</td>
<td>11.2/16.7</td>
<td>20.5/36.9</td>
<td>464/548</td>
<td>14.9/13.8</td>
</tr>
</tbody>
</table>

Proposed dataset, MWRAE as evaluation metric.

UCSD dataset [1], MAE and MWAE as evaluation metric.

<table>
<thead>
<tr>
<th>Method</th>
<th>Left</th>
<th>Right</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed CNN</td>
<td>0.6040</td>
<td>0.7281</td>
</tr>
</tbody>
</table>

References
1. Introduction

- Study of Visual Question Answering has led to the exploration of complex models designed to do multi-modal ‘reasoning’ and ‘attention’.
- We show that a basic approach designed to exploit answer bias performs well despite being much simpler:
  - reaches state-of-the-art performance on Visual7W Telling.
  - performs competitively on VQA Real multiple choice.
- Surprisingly, the model does ‘well’ even when missing the image or question.

2. Task

Visual Question Answering involves producing the correct text answers given a text question about an image. We consider the multiple-choice subtask, in which a smaller set of candidate answers is provided at test time.

3. MLP

![Diagram of MLP architecture]

Our model predicts correctness of an Image-Question-Answer triplet.

- **Loss**: We minimize the binary logistic loss of predicting triplet correctness: 
  \[ L(x, y) = -y \log f(x) + (1 - y) \log(1 - f(x)) \]

- **Optimization**: We fit model parameters with SGD, using 25 examples per batch, sampling 2 neg. for each pos., for 300 epochs.

- **Representation**:
  - **Images**: penultimate layer of a ResNet-101 trained on Imagenet.
  - **Text**: BoW (average) of pre-trained word2vec embeddings.
  - All features are off-the-shelf, not finetuned, and \( l_2 \) normalized.

4. Comparison with the State-of-the-Art

<table>
<thead>
<tr>
<th>Method</th>
<th>Where</th>
<th>Who</th>
<th>Why</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>69.0</td>
<td>54.1</td>
<td>51.3</td>
<td>52.1</td>
</tr>
<tr>
<td>LSTM+</td>
<td>71.0</td>
<td>55.5</td>
<td>50.8</td>
<td>53.5</td>
</tr>
<tr>
<td>MCB + Att [21]</td>
<td>60.3</td>
<td>70.4</td>
<td>69.2</td>
<td>62.2</td>
</tr>
</tbody>
</table>

5. Error Analysis

6. Exploiting Answer Similarity

The model can exploit answer similarity, NN by BoW cosine similarity:

7. Conclusion

Modeling the answer helps for VQA multiple-choice.
Are the evaluation protocols of current VQA tasks adequate?
Are current, more complex VQA models learning what we think they are?

For more information, contact {ajabri, ajoulin, lvdmaaten}@fb.com.
A Continuous Optimization Approach for Efficient and Accurate Scene Flow

Zhaoyang Lv[1], Chris Beall[1], Pablo F. Alcantarilla[3], Fuxin Li[4], Zsolt Kira[2], Frank Dellaert[1]

Introduction

**Goal:** Solve the dense 3D scene flow problem from a pair of stereo imagery.

**Contributions:** We propose a purely continuous formulation that:
1. shows that faster optimization is achievable by rethinking the geometry and motion estimation problem as a non-linear least-square problem;
2. with small over-segmented planes, optimizing on discrete segmentation is unnecessary, compared with [1][2][3];
3. nonlinearity can be overcome by our robust initialization method.
4. achieves the state-of-art accuracy, while being much faster than top competitors, e.g. [1][2][3].

Overview

**Occlusion Error-vs-time on KITTI Benchmark**

<table>
<thead>
<tr>
<th>KITTI Scene Flow 2015</th>
<th>KITTI Optical Flow 2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running time (milliseconds)</td>
<td>Running time (milliseconds)</td>
</tr>
<tr>
<td>Scene flow error (%)</td>
<td>Optical flow error (%)</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>0.7</td>
<td>0.8</td>
</tr>
<tr>
<td>0.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Our method is highlighted as **green**, which achieves top performance (low-left corner), both in accuracy and computation speed.

Conclusion

- We believe estimating geometry in a continuous domain is sufficient and more efficient, instead of jointly estimating segmentation in discrete feature space and geometry in Euclidean space.
- Initialization for the non-linear problem is important. Our multi-hypotheses RANSAC method, and step-by-step optimization, serve as important initialization to approach the final global optimal results.

References

Improving Multi-frame Data Association with Sparse Representations for Robust Near-online Multi-object Tracking

Loïc Fagot-Bouquet, Romaric Audiger, Yoann Dhome, Frédéric Lerasle

1. Context

Multi-object Tracking by detection
Sliding window-based approach

Classification with collaborative representations

Motivations:
- Reasoning on several frames improves performances with only a slight latency.
- Complex appearance models help distinguishing targets.
- Sparse representations widely used in Single Object Tracking and a few online Multiple Object Tracking approaches.

Main contributions:
- Energy based on sparse representations for the multi-frame data association.
- Penalty \( l_{\infty,1} \) norm more suited to the case of Multi-object Tracking.

2. Framework overview

Step 1: Sparse representations computation
Step 2: Optimization of energy \( E \) by MCMC sampling
Step 3: Trajectory update at the beginning of the sliding window

Superpixel symbols:
- \( D \): dictionary
- \( D_r \): dictionary composed by elements of track \( \tau \)
- \( \alpha_{d,I} \): sparse representation of detection \( d \)
- \( \alpha_{y_d} \): residual reconstruction over \( \tau \)
- \( \Delta \): trajectory update

3. Proposed appearance model

Main idea:
Promotes configurations consistent with the sparse representations of the detections.

\[ App(C) = \sum_{\tau \in C} \sum_{d \in \tau} \left\| y_d - D_r \alpha_{y_d} \right\|_2 \]

- \( y_d \): feature vector of detection \( d \)
- \( D_r \): dictionary composed by elements of track \( \tau \)
- \( \alpha_{y_d} \): sparse representation of detection \( d \)
- \( \alpha_{y_d} = \arg \min_{\alpha} \left\| y_d - D \alpha \right\|_2 + \lambda \| \alpha \|_1 \)
- \( \alpha_{y_d} \): residual reconstruction over \( \tau \)
- \( \left\| y_d - D_r \alpha_{y_d} \right\|_2 \): residual reconstruction error

4. Structured sparsity

\[ \alpha_i = \arg \min_{\alpha} \frac{1}{2} \left\| y_i - D \alpha \right\|_2^2 + \lambda \| \alpha \|_1 \]

Usual penalties \( \lambda \):
- \( l_1 \) norm: induces sparsity among all detections.
- \( l_{\infty,1} \) (or \( l_{1,2} \) group norm) induces sparsity among groups of detections.

Proposed penalty:
\[ \| \alpha \|_{\infty,1} = \max_{i=1,\ldots,N} w_i \| \alpha_i \|_1 \]

- Promotes representations where all groups are involved with a few elements inside each group.
- Frames as groups: favors to represent each detection with all the detections of the same target.

Optimization:
- Exact optimization using proximal gradient descent methods (PSTA) with active sets.
- Proximal operators for a weighted \( l_{\infty,1} \) norm efficiently evaluated with projections on the unit ball of the related dual norm.

5. Results

Evaluation on the MOTChallenge:
- Comparable results in MOTA with the majority of recent approaches.
- Best results in IDS and FAP, second best FM.
- 7.5 fps (4 cores CPU).

MOT16:
- Competitive with most of recent methods.

6. Conclusion

- New energy formulation exploiting sparse representations for multi-frame data association.
- \( l_{\infty,1} \)-based representations improve results compared to \( l_1 \)-based representations.
- Robust tracking in terms of identity switches and track fragmentation, comparing well with state-of-the-art approaches.
- Future work: extension to local features and joint sparse representation of all targets.

Contact: loic.fagot-bouquet@cea.fr
Saliency Detection via Combining Region-Level and Pixel-Level Predictions with CNNs

Youbao Tang, Xiangqian Wu

Email: tangyoubao@hit.edu.cn, xqwu@hit.edu.cn

INTRODUCTION

Visual saliency detection aims to highlight the most important object regions in an image.

Numerous image processing applications incorporate the visual saliency to improve their performance:
- image segmentation, cropping, object detection, image retrieval, etc.

This paper proposes a novel saliency detection method by combining region-level saliency estimation and pixel-level saliency prediction with CNNs (denoted as CRPSD).

Extensive quantitative and qualitative experiments on four public benchmark datasets demonstrate that the proposed method greatly outperforms the state-of-the-art saliency detection approaches.

METHODOLOGY

The Framework of the Proposed Method:

- Including three stages: pixel-level saliency prediction, region-level saliency estimation, and salient map fusion.

Region-level Saliency Estimation:
- Adaptive region generation
- Superpixel segmentation
- Graph-based agglomerative clustering
- The superpixels which are adjacent and have similar colors are clustered.
- Different images are segmented into different number of regions.

Pixel-level Saliency Prediction:
- Based on VGGNet-16.
- The outputs of the last two blocks are combined for multi-scale feature learning.
- Using cross-entropy loss function to compute the error:

\[ L(W) = - \frac{1}{N} \sum_{i=1}^{N} \log P(y_i = 1 | X, W) - (1 - \alpha) \frac{1}{M} \sum_{i=1}^{M} \log P(y_i = 0 | X, W) \]

Saliency Map Fusion:
- The region-level salient map and the pixel-level salient map are computed by using different information of images.
- They are complementary and can be fused to further improve the performance.

EXPERIMENTS

Datasets:
- SED, ECSSD, PASCAL-S, and HKU-4S.

Evaluation criteria:
- precision-recall (PR) curves, F-measure (F), weighted F-measure (wF), mean absolute error (MAE).

Compared with seventeen existing state-of-the-art saliency detection approaches.

Table 1. The wF and MAE of different saliency detection method on different test datasets [incl. b, k, g, and r] indicate that CRPSD performs the best among all methods.

Table 2. The wF of baseline and our methods on all test datasets.

Table 3. The mean shuffled-AUC of different fixation prediction methods on two saliency detection approaches.
**Introduction**

- **Overview:** Creating realistic textured 3D models of building exteriors using rental ads and street views.

- **Input:**
  - Floorplan of the house \( F \) and street view images of the house \( I \)

- **Output:**
  - Accurate location of the house
  - Vertical house dimensions such as floor, door, and window heights
  - Realistic 3D house model

- **Our Contributions:**
  - A novel application and its effective solution
  - A multi-purpose dataset: the SydneyHouse dataset

**Method**

- **Initial location:** By placing the floorplan onto the map. However, this is not consistent with streetview observations.

- **Parameterization:** The building’s floorplan allows for efficient parametrization: location \( x, y \), camera height \( h \), floor height \( f \), door height \( d \), and window heights \( a_l, a_u \).

- **Energy formulation:**
  \[ E(y; I, F) = E_{edge}(y; I, F) + E_{obj}(y; I, F) + E_{seg}(y; I, F) \]
  \[ + E_{sal}(y; I, F) + E_{app}(y; I, F) \]
  - \( E_{edge} \): Detection of vertical and horizontal edges (b). We define energy via a distance transform to the house.
  - \( E_{obj} \): Detection of windows and doors (c). \( E_{seg} \): Semantic segmentation (f). We define energies via overlap with the house.
  - \( E_{sal} \): Saliency (e). \( E_{app} \): Overall color appearance (d). We define energies via score of the house.

- **Learning and inference:** We learn the weights using structured-SVM and use exhaustive search for inference. This is efficient since we exploit integral geometry (generalization of integral images to 3D).

**SydneyHouse Dataset**

- We collected 174 random houses in Sydney, with floorplan, map, street views, and ground truth house configuration.
- We also processed sub-datasets for facade parsing and dense point correspondence, which are derived from our main dataset.

**Experiments**

- **Quantitative results:** Location error \( xy \) in m, IOU, and vertical height errors \( h, f, d, a_l, a_u \) in cm.

<table>
<thead>
<tr>
<th>range/m</th>
<th>gt%</th>
<th>time</th>
<th>( xy )</th>
<th>( h )</th>
<th>( f )</th>
<th>( d )</th>
<th>( a_l )</th>
<th>( a_u )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours</td>
<td>6.22</td>
<td>3.72</td>
<td>36.1</td>
<td>47.0</td>
<td>41.2</td>
<td>41.7</td>
<td>28.4</td>
<td></td>
</tr>
<tr>
<td>box-reg</td>
<td>19.0%</td>
<td>6.22</td>
<td>58.1</td>
<td>41.7</td>
<td>41.2</td>
<td>28.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>google</td>
<td>102.8</td>
<td>49.7</td>
<td>43.1</td>
<td>36.9</td>
<td>33.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ours</td>
<td>19.0%</td>
<td>6.22</td>
<td>49.7</td>
<td>43.1</td>
<td>36.9</td>
<td>33.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Impact of search range:**

<table>
<thead>
<tr>
<th>quant. thres./m</th>
<th>time</th>
<th>( xy )</th>
<th>( h )</th>
<th>( f )</th>
<th>( a_l )</th>
<th>( a_u )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>19.05</td>
<td>2.62</td>
<td>49.7</td>
<td>43.1</td>
<td>36.9</td>
<td>33.6</td>
</tr>
<tr>
<td>0.35</td>
<td>12.58</td>
<td>2.81</td>
<td>51.5</td>
<td>34.0</td>
<td>26.8</td>
<td>31.3</td>
</tr>
</tbody>
</table>

- **Impact of height quantization:**

<table>
<thead>
<tr>
<th>range/m</th>
<th>time</th>
<th>( xy )</th>
<th>( h )</th>
<th>( f )</th>
<th>( a_l )</th>
<th>( a_u )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours</td>
<td>6.22</td>
<td>3.72</td>
<td>36.1</td>
<td>47.0</td>
<td>41.2</td>
<td>41.7</td>
</tr>
<tr>
<td>box-reg</td>
<td>19.0%</td>
<td>6.22</td>
<td>58.1</td>
<td>41.7</td>
<td>41.2</td>
<td>28.4</td>
</tr>
<tr>
<td>google</td>
<td>102.8</td>
<td>49.7</td>
<td>43.1</td>
<td>36.9</td>
<td>33.6</td>
<td></td>
</tr>
<tr>
<td>ours</td>
<td>19.0%</td>
<td>6.22</td>
<td>49.7</td>
<td>43.1</td>
<td>36.9</td>
<td>33.6</td>
</tr>
</tbody>
</table>

- **Timing:** Inference runs 19.05s per house without parallelization.
Superpixel-based Two-view Deterministic Fitting for Multiple-structure Data

Guobao Xiao¹, Hanzi Wang¹, Yan Yan¹, David Suter²
¹Xiamen University, {guobaoxiao, hanzi.wang, yanyan}@xmu.edu.cn
²The University of Adelaide, david.suter@adelaide.edu.au

● Overview

✓ Superpixels are introduced for deterministic model fitting.
✓ A deterministic sampling algorithm is proposed to exploit the grouping cues of superpixels and the corresponding keypoint matching information.
✓ A novel model selection algorithm is developed to find all model instances in data.

Fig.1. Overview of the proposed SDF method for homography estimation. (a) An image pair with keypoint correspondences. (b) Superpixel generation. (c) The procedure of the proposed method. (d) The fitting result.

● Hypotheses generation

Partition keypoint matches into a set of groups based on the superpixels.
Extend some groups by combining small groups within a limited region to avoid degeneration in sampled subsets.
Sample significant data points in each extended group according to matching scores to generate model hypotheses.

Fig.2. An example illustrates that features within a superpixel have a high possibility of belonging to the same structure.

● Model selection

Generate model hypotheses
Select the significant hypotheses as the estimated model instance
Remove the redundant model hypotheses according to the estimated model instance

Fig.3. The main steps of the proposed model selection framework.

● Experiments

Compared with the fitting methods with randomized nature, SDF is tractable and can deterministically provide consistent solutions for model fitting.
Compared with several other deterministic methods (e.g., BnB and Astar), SDF is much faster, and it can achieve promising performance on image pairs with both single-structure and multiple-structure data.

Hypotheses generation

Fig.2. An example illustrates that features within a superpixel have a high possibility of belonging to the same structure.

● Conclusions

✓ Compared with the fitting methods with randomized nature, SDF is tractable and can deterministically provide consistent solutions for model fitting.
✓ Compared with several other deterministic methods (e.g., BnB and Astar), SDF is much faster, and it can achieve promising performance on image pairs with both single-structure and multiple-structure data.

● References

**Motivation**

Fisher vector consists of many components with each corresponding to a Gaussian model.

We tend to select the Domain Invariant Components (DIC) of Fisher vectors by assigning higher weights to the components with smaller Maximum Mean Discrepancy (MMD). We learn the regression matrix and weights simultaneously.

For cross-domain (i.e., dataset) images, DIC means object regions with low intra-class variance. For instance, the handle is DIC of beer mug because they are less affected by the pattern and shape of mug.

For cross-domain (i.e., view) videos, DIC means view-invariant human action trajectories, which represent the movements visible to various views.

**Formulation**

For ease of optimization, we divide Fisher vector into category-specific segments and introduce an intermediate variable S, which is enforced to be close to R.
Deep Robust Encoder through Locality Preserving Low-Rank Dictionary

Zhengming Ding¹, Ming Shao¹, and Yun Fu¹,²

¹Department of Electrical & Computer Engineering, Northeastern University, Boston, USA
²College of Computer & Information Science, Northeastern University, Boston, USA

Problem

Conventional auto-encoder and its variants usually involve additive noises (e.g., Gaussian, masking) for training data to learn robust features, which, however, did not consider the already corrupted data.

Objective

\[
W_1, b_1, W_2, b_2, D = \min \sum_{i=1}^{m} \| D(x) - f_2(f_1(x_i)) \|^2 + \lambda \| D(x) \|^2
\]

Deep Architecture

- Considering the learning objective as a basic building block, we can train a more discriminant deep model.
- In this paper, we jointly learn deep structure by further adding each encoding layer with the graph regularizer.

Contributions

- A low-rank dictionary and deep AE are jointly optimized, which can progressively denoise the already corrupted features.
- The newly designed loss function penalizes the corruptions or distortions, meanwhile ensures that the reconstruction is noise free.
- Graph regularizers are developed to guide feature learning in each encoding layer to preserve more geometric structures within the data.

Reference

1. Introduction

1.1 Motivation

• Prior bottom-up methods predict saliency maps by combining heuristic saliency cues, which may be unreliable and lead to error outputs.
• Performance of traditional label propagation methods rely on the quality of saliency seeds as well as the propagation ability.

1.2 Solution

• Pattern mining based saliency seeds selection method
  The pattern mining based method is capable of high-level understanding and recognizes discriminative and representative saliency patterns, thus output more accurate saliency seeds.
• ERW based seeds propagation method
  The proposed External Random Walk (ERW) algorithm in one hand ensures the diffusion of seeds information to more distant areas and in other hand makes full use of known saliency knowledge.

2. Overview

• Saliency seeds selection
  Given initial saliency map generated by existing model, discriminative and representative saliency seeds are detected based on pattern mining method.
• Seeds label propagation
  The label of saliency seeds are propagated to other image regions using the proposed ERW algorithm.

3. Algorithm

3.1 Saliency seeds selection

• Finding saliency patterns
  - Each instance in sample pool is described by its visual word index, the visual word indexes of its K nearest neighbors and its class label (pos or neg).
  - Explore item sets \( A \) which frequently occur in positive samples but rarely in negative samples using the following criteria:
    \[
    \text{support}(A) > t_1, \quad \text{confidence}(A \rightarrow \text{pos}) > t_2,
    \]
• Detecting saliency seeds

3.2 Saliency Propagation

• ERW algorithm
  \[
  \begin{align*}
  \hat{f}^* &= \arg \min f \quad \frac{1}{2} \sum_{j} w_j (f_i - f_j)^2 + \frac{\alpha}{2} \sum_{j \in N(i)} w_{ij} (f_i - f_j)^2 + \frac{\beta}{2} \sum_{j \in N(i)} (f_i - y_j)^2, \\
  \text{s.t.} f_i &= 1, \quad \forall y_j \in \mathcal{L}, \\
  \end{align*}
  \]

• Solution
  \[
  \hat{f}_n = M^{-1}_{wn} \left(-M_{wn} f_i + \beta y_n \right)
  \]
  \[
  M = L + \alpha \Lambda^2 + \beta I
  \]

4. Results

(a) MSRA
(b) SOD
(c) ECSSD
(d) PASCAL-S
Streaming Video Segmentation via Short-Term Hierarchical Segmentation and Frame-by-Frame Markov Random Field Optimization

Korea University
MCL (http://mcl.korea.ac.kr/)

Won-Dong Jang and Chang-Su Kim

ECCV’16

Introduction

• An online video segmentation algorithm
  • Divide a video into volumetric segments

• Overview of the proposed algorithm

Proposed Algorithm

• Feature extraction for each frame
  • Estimate both forward and backward optical flows
  • Over-segment each frame into superpixels (Mean-shift)
  • Color feature
    • LAB histogram (20 bins for each channel)
    • LAB and RGB BoW histogram (300 and 300 words)

• Motion feature
  • Forward motion BoW histogram (100 words)
  • Backward motion BoW histogram (100 words)

• Boundary feature
  • Image segmentation using various parameters

• Short-Term Hierarchical Segmentation (STHS)
  • Perform initial segmentation of frame \( \tau \) using frames from \( \tau - \alpha \) to \( \tau + \alpha \), where \( \alpha = 7 \)

• Spatial agglomerative clustering
  • Construct a graph \( G^{(\tau)} = (V^{(\tau)}, E^{(\tau)}) \) for frame \( \tau \)
  • Set superpixels as nodes
  • Connect an edge between superpixels that share a boundary
  • Initially regard superpixels as individual clusters
  • Measure distances between clusters
    \[
    d(c_i, c_j) = \begin{cases} 
    d_{\text{opt}}(x_i, x_j) & \text{if } e_{ij} \in E^{(\tau)} \\
    \infty & \text{otherwise}
    \end{cases}
    \]

  • Merge the two clusters \( c_i \) and \( c_j \) that yield the minimum distance

  • Update the distance between the new cluster \( c_k \) and an existing cluster \( c_k \)
    \[
    d(c_k, c_k) = \min \{ d(c_k, c_i), d(c_k, c_j) \}
    \]

  • Terminate the merging when the minimum distance is higher than a threshold \( \gamma \) — Controls the segmentation granularity

• Temporal graph matching
  • Perform the matching between two frames sequentially
  • Construct a bipartite graph \( G^{(\tau+1)} = (V^{(\tau)}, U^{(\tau+1)}, E^{(\tau+1)}) \)
  • Set clusters as nodes
  • Connect an edge between clusters using optical flows
  • Compute an affinity weight for each edge
    \[
    \eta^{(\tau+1)}(c_i^{(\tau)}, c_j^{(\tau+1)}) = \begin{cases} 
    \eta_{\text{sim}}(c_i^{(\tau)}, c_j^{(\tau+1)}) & \text{if } e_{ij} \in E^{(\tau+1)} \\
    0 & \text{otherwise}
    \end{cases}
    \]

  • A similarity function between the two clusters
    \[
    \eta_{\text{sim}}(c_i^{(\tau)}, c_j^{(\tau+1)}) = \mu_{\text{sim}}(c_i^{(\tau)}, c_j^{(\tau+1)}) \times \mu_{\text{gr}}(c_i^{(\tau)}, c_j^{(\tau+1)})
    \]

• MRF optimization
  • Define the MRF energy function using the graph \( G^{(\tau)} = (V^{(\tau)}, E^{(\tau)}) \)
    \[
    \mathcal{E}(y^{(\tau)}) = \sum_{v \in V^{(\tau)}} \psi(x_i^{(\tau)}, y_i^{(\tau)}) + \sum_{(j,k) \in E^{(\tau)}} \phi(x_i^{(\tau)}, x_j^{(\tau)}, y_i^{(\tau)}, y_j^{(\tau)})
    \]

  • Unary cost
    \[
    \psi(x_i^{(\tau)}, y_i^{(\tau)}) = -\log p(y_i^{(\tau)} | x_i^{(\tau)}) = -\log \left( \sum_k \theta_i(y_i^{(\tau)} | x_i^{(\tau)}, k) \right)
    \]

  • Pairwise cost
    \[
    \phi(x_i^{(\tau)}, x_j^{(\tau)}, y_i^{(\tau)}, y_j^{(\tau)}) = \begin{cases} 
    \exp(-d_{\text{sim}}(x_i^{(\tau)}, x_j^{(\tau)})) & \text{if } y_i^{(\tau)} \neq y_j^{(\tau)} \\
    0 & \text{otherwise}
    \end{cases}
    \]

  • Minimize the MRF energy function using the graph-cut algorithm

Experimental Results

• Demo video

• VSB100 dataset
• Performance metrics
• Volume precision-recall (VPR)

Benchmark table

<table>
<thead>
<tr>
<th>BPR</th>
<th>VPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ODS</td>
<td>OSS</td>
</tr>
<tr>
<td>Human</td>
<td>0.81</td>
</tr>
<tr>
<td>A. Offline segmentation algorithms</td>
<td></td>
</tr>
<tr>
<td>[1]</td>
<td>0.64</td>
</tr>
<tr>
<td>[2]</td>
<td>0.63</td>
</tr>
<tr>
<td>[3]</td>
<td>0.38</td>
</tr>
<tr>
<td>[4]</td>
<td>0.63</td>
</tr>
<tr>
<td>Ours</td>
<td>0.63</td>
</tr>
<tr>
<td>Ours (Faster)</td>
<td>0.63</td>
</tr>
</tbody>
</table>


• Segmentation granularity

Input frame Ground-truth \( \gamma = 0.1 \) \( \gamma = 0.2 \) \( \gamma = 0.4 \)
Learning to Learn: Model Regression Networks for Easy Small Sample Learning

Yu-Xiong Wang and Martial Hebert
Email: {yuexiongw, hebert}@cs.cmu.edu

**Motivation**

- **Inter-class knowledge**: A generic transformation $T$: small-sample models $w^0 \rightarrow$ large-sample models $w^*$
- **Recognition of novel categories from few examples**: Predict the target models $w$ by transferring $T$

**Learning Model Transformation**

- $T$ learns predictive structures in the model space
  - Discriminative representation of natural intra-class variability: Sparse samples $\rightarrow$ a category cluster
  - Duality perspective: Feature $\leftrightarrow$ classifier spaces
  - Alternative parametric way of model distillation
- $T$ can be learned
  - On a large collection of model pairs $\{ (w^0_j, w^*_j) \}$
  - By a high-capacity regression function $T (w^0_j, \Theta)$

**Evaluation**

- **Fine-grained, action & scene classification**
- **Baselines**: Models learned from scratch & model parameter transfer & fine-tuning

**One-Shot Domain Adaptation**

<table>
<thead>
<tr>
<th>Prior knowledge</th>
<th>Method</th>
<th>Acc (%)</th>
<th>Prior knowledge</th>
<th>Method</th>
<th>Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
<td>SVM (target only)</td>
<td>62.28</td>
<td>Feature</td>
<td>GPK</td>
<td>65.16</td>
</tr>
<tr>
<td>Data</td>
<td>SVM (source only)</td>
<td>53.51</td>
<td>SA</td>
<td>59.30</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SVM (source and target)</td>
<td>56.68</td>
<td>Daumé III</td>
<td>59.21</td>
<td></td>
</tr>
<tr>
<td>Model parameter</td>
<td>PMT</td>
<td>66.30</td>
<td>MMDT</td>
<td>59.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Late fusion (Max)</td>
<td>59.59</td>
<td>Joint</td>
<td>Fine-tuning</td>
<td>61.13</td>
</tr>
<tr>
<td></td>
<td>Late fusion (Lin. Int. Avg)</td>
<td>60.64</td>
<td>Model transformation</td>
<td>Regression network (Ours)</td>
<td>68.47</td>
</tr>
</tbody>
</table>

**Prior Knowledge**

- SVM (target only)
- SVM (source only)
- SVM (source and target)
- PMT
- Late fusion (Max)
- Late fusion (Lin. Int. Avg)
- Feature
- GPK
- SA
- Daumé III
- MMDT
- Joint
- Fine-tuning
- Regression network (Ours)
1. Introduction

- Problems of Three-Dimensional Geometry Recovery:
  - It is difficult to reconstruct objects with more complex materials such as translucency, subsurface scattering, and interreflections.
  - The medium between the surface and the camera is treated as an unwanted nuisance (triangulation, time of flight, and shape-from-X).
  ⇒ We focus on light absorption in the infrared spectrum.
- Key Idea:
  These are 950nm images of pouring water into the cup.

2. Method

1. Water Absorption
   Beer-Lambert law:
   \[ I = I_0 e^{-\alpha l} \]
   - \( I \): light intensity after passing medium
   - \( I_0 \): light intensity before passing medium
   - \( \alpha \): absorption coefficient
   - \( l \): medium length
   - The water absorption curve in the range from 400nm to 1600nm (with 6mm depth)
   - The water clearly absorbs infrared light from 900nm to 1400nm.

2. Bispectral Shape from Water
   - The intensity of the light received are
     \[ I(\lambda_1) = I_0(\lambda_1) e^{-\alpha_1 l}, \]
     \[ I(\lambda_2) = I_0(\lambda_2) e^{-\alpha_2 l}. \]
   - Assuming \( s(\lambda_1) = s(\lambda_2) \),
   - the approximate depth can be recovered,
     \[ l = \frac{1}{2(\alpha(\lambda_2) - \alpha(\lambda_1))} \ln \frac{I(\lambda_1)}{I(\lambda_2)}. \]

3. Our assumptions and verification
   - Orthographic Camera
   - Parallel Incident Light Rays to the Object Surface
   - Flat Spectrum in a short range of NIR

3. System

- We built a co-axial bispectral imaging system for shape from water. The system uses co-axial cameras to simultaneously capture the scene in two wavelengths, recording bispectral image pairs at video-rate.

4. Experimental Results

1. Accuracy
   - We use planar plates with different materials for depth accuracy evaluation. We put the plates in water and measure the water depth by a ruler for ground truth. We vary the water depth from 10mm to 40mm.

2. Reconstruct Results
   - Shape recovery of objects with complex geometry, texture, and reflection properties.
Learning Dynamic Hierarchical Models for Anytime Scene Labeling

Buoy Liu, Xuming He
The Australian National University (ANU), Data61, CSIRO, Australia

Problem Setting

- Cost-aware scene parsing
- Learning a scene labeling model such that
  - Stop at any given budget
  - Achieve good performance
  - Flexible traded-off between accuracy and efficiency

Main Contribution

- Anytime scene labeling that incorporates the cost of feature computation and model inference
- Dynamic hierarchical models to achieve flexible trade-offs between efficiency and accuracy

Overview

- Build a family of hierarchical scene models
- Model selection as an Markov Decision Process

Model Formulation

- Model generation: Dynamic Hierarchical Model (DHM)
  - Hierarchical model: \(M'\)
    - Segmentation tree: set of leaf nodes \(B' = \{b_1, \ldots, b_M\}\)
    - Prediction on leaf nodes \(\mathcal{Q}' = \{q_{b_i}^{(k)}\}_{i=1}^{N}\)
  - Sequential model generation process:
    - Split-inherit: \(q_{b_i}^{(k+1)}(k) = q_{b_i}^{(k)}(k), \quad k \in \mathcal{Y}, \quad i \in \mathcal{G}^{t+1} \quad \text{if } q_i > \theta_i\)
    - Local belief update: \(q_{b_i}^{(k+1)}(k) \propto q_{b_i}^{(k)}(k) \exp(a_b [f_b^C(k)])\)
    - \(k \in \mathcal{Y}, \quad \mathcal{F}^C = \{x_a, q_{b_i}^{(k)}\}\)

- DHM generation as Markov Decision Process (MDP)
  - State \(s_t = (B', \mathcal{Q}', \mathcal{G}');\) active nodes \(\mathcal{G}^{t} \in \{0, 1\}^{N}\)
  - Action \(a_t = b_i \) or \(a_t = a_b;\) split-inherit or belief update
  - Transition \(T(s_{t+1}, s_t, a_t)\)
  - Reward \(r_t\) and discount factor \(\gamma_t\)
    - cost-sensitive & encouraging large improvement at each step
  - Reward function
    - Labeling loss: balance between segmentation quality and prediction accuracy
      \(L(s_t, Y_t) = -\sum_{i \in \mathcal{G}^t} w_{p_i} \log(p_{s_i}) - \alpha \sum_{i \in \mathcal{G}^t} w_{p_i} \log(p_{s_i})\)
    - Cost-sensitive reward:
      \(R(s_t, a_t) = \frac{1}{c(a_t)} \left[ L(s_t, Y_t) - L(s_{t+1}, Y_{t+1}) \right]\)
      \(c(a_t)\) computes image feature, region split, feature pooling and weak learner cost

- Our objective for anytime scene labeling
  - Policy of the MDP \(\pi(s) : \mathcal{S} \rightarrow \mathcal{A}\)
  - Find an optimal policy \(\pi^*\) to maximize the accumulative reward:
    \(V_{\pi}(s_0) = \sum_{t=0}^{\infty} \gamma^t R(s_t, \pi(s_t))\)
  
Policy learning

- Learning an optimal policy is challenging:
  - Large state space & discrete-continuous action space

- Our solution
  - Discretize the action space by generating action proposals: greedy method [1]
  - Adopt a linear function to approximate the policy
    \(Q(s_t, a_t) = \eta^T \phi(s_t, a_t),\)
    \(\phi(s_t, a_t)\) : meta-feature, including:
      - Statistics of region prediction
      - Feature from leave nodes
  - Use Q-learning to find a ‘local optimal’ linear policy

Scene labeling at test time

- While Test-time cost is not reached do
  - Compute \(\phi(s_{t+1}, a_t)\) for all \(a_t \in \mathcal{A}^t\);
  - if max \(\eta^T \phi(s_{t+1}, a_t) \leq 0\) then break;
  - end

Experiment details: CamVid, Stanford background and Siftflow
- Feature set: 9 different types of visual features (including CNNs features)
- Baselines: S-M: static-mycopic; RS: random selection; F-SM: DCRF + feature selection

Results on CamVid

- Anytime performance on CamVid

- Ablation study on CamVid

- Semantic segmentation results

Results on Stanford background and Siftflow


Acknowledgment

This project is funded by the Australian Government as represented by the Dept. of Communications and the ARC through the ICT Centre of Excellence program.
Semantic 3D Reconstruction of Heads

Fabio Maninchetta\textsuperscript{1}, Christian Häne\textsuperscript{2, *}, Bastien Jacquet\textsuperscript{3, *}, Amaël Delaunoy\textsuperscript{*}, and Marc Pollefeys\textsuperscript{1, 4}

\textsuperscript{1}ETH Zürich — \textsuperscript{2}UC Berkeley — \textsuperscript{3}Kitware — \textsuperscript{4}Microsoft — *Work done at ETH Zürich

Introduction
We propose a system that automatically reconstructs and semantically segments human heads from images taken in uncontrolled environments using a multi-label shape prior.

Formulation
- Labeling of a voxel space into \( L \) labels \( x^i \in [0,1] \) and \( \sum x^i = 1 \)
- Labels: free space, skin, hair, eyebrows, beard and clothing
- Unary cost contains depth information
- Smoothness cost depends on surface orientation and involved labels (contains shape prior and semantic information)

Optimization Problem
\[
E(x, T) &= \sum_{i \in \Omega} \left( \sum_{j \in \Omega} \phi(T, x^j) x^i + \frac{1}{T} \sum_{j \in \Omega} \phi(T, x^j - x^i) \right) \\
\text{s.t.} \quad x^i &= \sum_{j \in \Omega} (x^j)_{s_i}, \quad x^i = \sum_{j \in \Omega} (x^j)_{s_i}, \quad \sum x^i = 1, \quad x^i \geq 0, \quad \sum x^i \geq 0.
\]

Unary Term
- \( \phi(T, x^j) \): alignment between prior and input data
- \( \mu^j \): unary term cost, for label \( i \) at location \( s \)
- \( \phi(T, x^j - x^i) \): convex smoothness term at voxel \( s \) for labels \( i \) and \( j \)
- \( \omega_k \in \mathbb{R}^3 \): \( k \)-th canonical basis vector
- \( T \): alignment between prior and input data
- \( \gamma \): normalization factor for scale
- Convex in \( x \), not convex in \( T \) (optimized using alternating optimization)

Data Dependent Regularization Term
- Smoothness term \( \phi(T, x^j) \) in terms of a convex Wulff shape \( W^j \), approximated as intersection of half spaces
- \( d_{T,x} \): cost for surface with normal direction \( n \) between labels \( i \) and \( j \), corresponds to distance of half space boundary to origin
- We have \( \phi(T, x^j) = d_{T,x} \)
- \( d_{T,x} = -\log(P(n(T), x^j)) \) is approximated with
- \( d_{T,x} = -\log[P(n(T)x^j + n^T)] - \log[P(n^T)x^j] \)
- Normal based shape prior: locally, surface normals are similar between different heads
- General Wulff shape replaced by surrogate when all normals point in similar direction

Acknowledgment
This project is supported by Grant 16703.1 PFES-ES of CTI Switzerland and the Swiss National Science Foundation under Project Nr. 143422
Motivation

- Human attribute recognition: a fine-grained classification problem.
- Leverage contextual cues to make human attributes more recognizable:
  - Neighboring similar people are likely to have similar attributes.
  - People’s attributes are related to the scenes they appear in.

Deep Hierarchical Contexts

- Framework: Fast-RCNN + VGG16.
- Four attribute scoring branches using different cues:
  - Person bounding box.
  - Attribute-specific parts.
  - Human-centric context.
  - Scene-level context.

- Human-centric context: select nearest neighbor parts of other people on the deep feature pyramid (below left).
- Scene-level context: map scene classification score to attribute prior probability (below right).

Dataset: WIDER Attribute

- WIDER Attribute is a large-scale human attribute dataset with image event annotations. It contains 13789 images belonging to 30 scene categories, and 57524 human bounding boxes each annotated with 14 attributes.

Results

- Mean AP (%) on 3 datasets: Berkeley Attribute of People, HAT and WIDER Attribute (B: baseline, HC: human-centric context, SC: scene-level context):

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Images</th>
<th>Human Bounding Boxes</th>
<th>Bboxes Per Image</th>
<th>Attributes</th>
<th>Scenes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Berkeley</td>
<td>8035</td>
<td>17628</td>
<td>2.2</td>
<td>9</td>
<td>-</td>
</tr>
<tr>
<td>HAT</td>
<td>9344</td>
<td>19872</td>
<td>2.1</td>
<td>27</td>
<td>-</td>
</tr>
<tr>
<td>PARSE-27k</td>
<td>20999</td>
<td>27454</td>
<td>1.3</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>WIDER Attribute</td>
<td>13789</td>
<td>57524</td>
<td>4.2</td>
<td>14</td>
<td>30</td>
</tr>
</tbody>
</table>

- Illustration of image scenes:

Visualizations of attribute recognition results and scene prediction:

Conclusion

- We introduced a large-scale WIDER Attribute dataset with rich human attribute and event class annotations.
- Project page:
  http://mmlab.ie.cuhk.edu.hk/projects/WIDERAttribute.html
Introduction

- We address the person re-identification problem by exploiting a globally feature representation from a sequence of tracked human regions/patches.
- We show that a progressive fusion framework based on LSTM aggregates the frame-wise human region representation and yields a sequence level feature representation.
- Experimental results on two person re-identification benchmarks demonstrate that the proposed method performs favorably against state-of-the-art person re-identification methods.

Recurrent Feature Aggregation Framework

- LBP and Color features are first extracted from rectangular image patches, and then concatenated as frame level representation.
- This representation and the previous LSTM outputs are input to the current LSTM node.
- At each time stamp $t$, given the input $x_t$ and the previous hidden state $h_{t-1}$, we update the LSTM network as follows:

$$i_t = \sigma(W_x x_t + U_h h_{t-1} + V_c c_{t-1} + b_i)$$ (1)

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1} + b_f)$$ (2)

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_c x_t + U_c h_{t-1} + b_c)$$ (3)

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t + b_o)$$ (4)

$$h_t = o_t \cdot \tanh(c_t)$$ (5)

- The output of the LSTM hidden state $h_t$ is further connected to a softmax layer. The output of the $N$-way softmax is the prediction of the probability distribution over $N$ different identities.

$$y_i = \frac{\exp(y_i)}{\sum_{k=1}^N \exp(y_k)}$$ (6)

- The network is learned by minimizing $-\log y_k$, where $k$ is the index of the true label for a given input. Stochastic gradient descent is used with gradients calculated by back-propagation.

Experimental Results

- We address the person re-identification problem by exploiting a globally feature representation from a sequence of tracked human regions/patches.
- We show that a progressive fusion framework based on LSTM aggregates the frame-wise human region representation and yields a sequence level feature representation.
- Experimental results on two person re-identification benchmarks demonstrate that the proposed method performs favorably against state-of-the-art person re-identification methods.

Table 1. Performance of different methods on Color&LBP feature

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iLIDS-VID</th>
<th>PRID 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color&amp;LBP + RSVM</td>
<td>23.2 44.2 54.1</td>
<td>68.8 34.3 56.0 65.5 77.3</td>
</tr>
<tr>
<td>Color&amp;LBP + DTW [2]</td>
<td>9.3 21.7 29.5</td>
<td>43 14.6 33 42.6 47.8</td>
</tr>
<tr>
<td>Color&amp;LBP + DVR [3]</td>
<td>34.5 56.7 67.5</td>
<td>77.5 37.6 63.9 75.3 89.4</td>
</tr>
<tr>
<td>Color&amp;LBP + RFA-Net + Cosine</td>
<td>44.5 71.9 82.0</td>
<td>90.1 54.9 84.2 93.7 98.4</td>
</tr>
<tr>
<td>Color&amp;LBP + RFA-Net + RSVM</td>
<td>49.3 76.8 85.3</td>
<td>90.0 58.2 85.8 93.4 97.9</td>
</tr>
</tbody>
</table>

Table 2. Performance of our method in existence of noises

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iLIDS-VID</th>
<th>PRID 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise Level: 5%</td>
<td>49.3 76.8 86.3</td>
<td>90.0 58.2 85.8 93.4 97.9</td>
</tr>
<tr>
<td>Noise Level: 10%</td>
<td>43.4 70.6 81.5</td>
<td>88.9 52.3 83.2 91.4 97.5</td>
</tr>
<tr>
<td>Noise Level: 30%</td>
<td>40.0 67.4 77.5</td>
<td>87.0 51.4 81.1 90.5 96.9</td>
</tr>
<tr>
<td>Noise Level: 50%</td>
<td>29.8 60.5 71.9</td>
<td>81.5 44.7 75.2 85.6 95.5</td>
</tr>
</tbody>
</table>

Table 3. Performance of our method compared against state-of-the-art methods.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>iLIDS-VID</th>
<th>PRID 2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>STFV3D [4]</td>
<td>37.0 64.3 77.0</td>
<td>86.9 21.6 46.4 58.3 73.8</td>
</tr>
<tr>
<td>STFV3D+KISSME</td>
<td>44.3 71.7 83.7</td>
<td>91.7 64.1 87.3 89.9 92.0</td>
</tr>
<tr>
<td>Color&amp;LBP + Our + Cosine</td>
<td>44.5 71.9 82.0</td>
<td>90.1 54.9 84.2 93.7 98.4</td>
</tr>
<tr>
<td>Color&amp;LBP + Our + RSVM</td>
<td>49.3 76.8 85.3</td>
<td>90.0 58.2 85.8 93.4 97.9</td>
</tr>
</tbody>
</table>

References


Motivation

- Many important problems require solving low rank SDPs with PSD constraint matrices.
- SDPs are commonly used in binary valued labeling problems (BQP) using semidefinite relaxation (SDR).
- BQP, SDR transforms non-convex rank-1 constraint to PSD matrices with all-ones diagonals.
- Existing solvers are problem specific or slow.
- Address challenges of memory and time; more general solver.

Contribution

- A novel framework to approximately solve SDPs with PSD constraint matrices efficiently.
- Effective initialization.
- 4x to 35x speedups with state-of-the-art results.

Overview

General SDPs

\[
\text{minimize} \quad \sum_{i \in E} \langle Q_i, X \rangle \\
\text{subject to} \quad \|Q_i - L_i X\|_F^2 = b_i, \quad \forall i \in \{B \cup E\}
\]

Specialized Solvers

- Interior Point Method
- Convex Algorithm
  - Spectral sub gradient method
  - Dual gradient descent
- Non-convex Algorithm
- Low rank approximation
- Augmented Lagrangian

Biconvex Relaxation (BCR) Framework

\[
\text{minimize} \quad \sum_{i \in \{B \cup E\}} \langle Q_i, X \rangle \\
\text{subject to} \quad \|Q_i - L_i X\|_F^2 = b_i, \quad \forall i \in \{B \cup E\}.
\]

Non-convex formulation

\[
\text{minimize} \quad \frac{1}{2} \sum_{i \in \{B \cup E\}} \|Q_i - L_i X\|_F^2 \\
\text{subject to} \quad \|Q_i\|_F^2 = b_i, \quad \forall i \in \{B \cup E\},
\]

Concave penalty counters relaxed equality constraints

where $\alpha > \beta > 0$ are relaxation parameters.

- Generally, $\alpha = 2\beta$ and $\beta = \|C\|_2$
- $\beta$ is chosen to match the curvature of objective with that of penalty term.
- The problem is biconvex when $C \in S^+_{X \times X}.$

Initialization

- If poorly initialized, Alternating Minimization algorithm may get trapped in local minima.
- Extend initialization algorithm previously used for phase retrieval problem [2, 4].
- Decompose $C$ into $C = U^T U$ and use $\tilde{X} = U X.$
- Then initialize $X_0 = \lambda V$, where $\lambda$ and $V$ are the $r$ leading eigenvalue and eigenvectors of $Z = \frac{1}{\|C\|} \sum_{i \in E} b_i^{1/2} \ll_i T \ll_i.$
- Initialize $X_0 = U^{-1} \tilde{X}_0.$

Optimization

- Alternating minimization Algorithm
  - Stage 1: Minimize w.r.t. $(Q_i),$ i.e., minimize quadratic objective.
  - Project $Q_i$ back into unit Frobenius-norm ball of radius $\sqrt{b_i}.$
  - Expansion-reprojection update, $Q_i \leftarrow X \leftarrow \min_{\|X\|_F \leq \sqrt{b_i}} \langle C + \alpha \sum_{i \in \{B \cup E\}} L_i^T L_i, \sum_{i \in \{B \cup E\}} L_i^T Q_i \rangle$
  - Stage 2: Minimize w.r.t. $X.$ Solves the least-squares problem,

Numerical Experiments

General Form Problems

Image Segmentation

- Berkeley dataset. The red and blue marker indicates annotated foreground and background super pixels.

Metric Learning on Manifolds

Image Co-segmentation

- Weizmann horses and MSRC datasets.

References

**Algorithm Overview**

**Motivation**
- Idea is to find what is “common” in a set of images

**Problem Definition**
- Given more than one image, perform image co-segmentation to obtain
  - objects with visually similar feature
  - objects may be different in size
  - multiple common objects, if present
  - exclude similar background

**Possible Approaches**
- Supervised co-segmentation [Batra et al., 2010]
- Graph cut based approach [Rother et al., 2006]
- Saliency based approaches [Rubinstein et al., 2013]

**Region Adjacency Graph (RAG)**
- Image superpixels as nodes
- Node attribute
  - Color mean in CIELab color space
  - Rotation invariant HoG feature
- Edge between adjacent nodes (superpixels)

**Standard Maximum Common Subgraph (MCS) Algorithm**
- Compute vertex product graph (VPG) from the input graph pair (RAGs)
  - node attribute (many-to-many matching, strict threshold)
  - edges (region adjacency constraints)
- Maximum clique in the VPG is the MCS [Koch 2001]
  - obtained from minimum vertex cover (MVC) of the complement of the VPG [Cormen2001]

**Flow of the Algorithm**

**Region Co-growing (RCG)**
- MCS outputs partially detect common objects
  - different size, pose of objects in natural images
- Use MCS outputs as seeds and simultaneously grow in both images and iterate
  - Feature similarity between a matched node in RAG1 and neighbors of matched nodes in RAG2 and vice-versa
  - relaxed threshold
- Append newly matched neighbors
- Measure of Zhu et al.[2014] to discard background superpixels

**Multi-image Co-segmentation**
- Perform multiple levels of pair wise co-segmentation

**Hierarchical Co-segmentation**
- Computational complexity
- Hierarchical image resizing
  - Compute MCS at the coarsest level
  - Map matched region co-ordinates to the coarsest level to finer levels
  - Perform RCG at the finer levels

**Results – Visual Comparison**

**Quantitative Comparison**

**Conclusions**
- Inexact MCS: feature similarity
- MCS stage: multiple objects
- RCG stage: different sized objects
- Hierarchical co-segmentation: process images of large size
Motivation and Idea:

End-to-end deep network to simultaneously rank and localize relative attributes.

Attribute: Smile

• Local representation of visual attributes often leads to better performance [Bourdev ICCV 11, Duan CVPR 12, Sandeep CVPR 14].
• Existing localization methods have two limitations:
 جوز require human supervision [Bourdev ICCV 11, Duan CVPR 12, Sandeep CVPR 14].
 جوز not optimized jointly to rank and localize attribute [Xiao & Lee ICCV 15].
• Key Idea: Jointly optimize relative attribute ranking and localization with an end-to-end deep network.

Approach:

For training, pair of images \( \{I_1, I_2\} \) and label denoting their relative ordering is given as input to Siamese network which consists of:

- Spatial Transformer Network (STN): Learn image transformation parameters \( (t_x, t_y, s) \) to choose most relevant image region for ranking.

\[
\begin{align*}
I_{out}^{(s)} &= I + (s \cdot t_x - t_y) \\
I_{out}^{(s)} &= I + (s \cdot t_x - t_y)
\end{align*}
\]

- Ranker Network (RN): Takes STN output and original image (for context) as input, and combine their features to generate attribute strength \( v \).

For testing, one branch of the Siamese network is used to predict attribute strength.

Network Architecture and Loss:

Ranking Loss:

- \( v_1 > v_2 \) when \( I_1 \) has higher attribute strength than \( I_2 \) \( L = 1 \).
- \( v_1 = v_2 \) when \( I_1 \) and \( I_2 \) have similar attribute strengths, \( L = 0.5 \).

\[
\text{Rank}_{\text{loss}}(I_1, I_2) = -L \cdot \log(P) - (1 - L) \cdot \log(1 - P),
\]

where \( P = \frac{e^{v_1 - v_2}}{1 + e^{v_1 - v_2}} \)

Spatial Transformer Loss: If output of STN goes outside of image boundaries:

\[
S_{\text{Loss}}(I) = (C_x - s \cdot t_x)^2 + (C_y - s \cdot t_y)^2,
\]

where \( (C_x, C_y) \) is image center

Stochastic gradient descent is used to optimize the loss.

Quantitative Results:

<table>
<thead>
<tr>
<th>LFW-10</th>
<th>BH</th>
<th>DH</th>
<th>EO</th>
<th>GL</th>
<th>ML</th>
<th>MO</th>
<th>S</th>
<th>VT</th>
<th>VF</th>
<th>Y</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>P &amp; G + CNN</td>
<td>78.10</td>
<td>83.09</td>
<td>71.43</td>
<td>68.73</td>
<td>95.40</td>
<td>65.77</td>
<td>63.84</td>
<td>66.46</td>
<td>81.25</td>
<td>72.07</td>
<td>74.61</td>
</tr>
<tr>
<td>Sandeep et al.</td>
<td>82.04</td>
<td>80.56</td>
<td>83.52</td>
<td>68.98</td>
<td>90.94</td>
<td>82.04</td>
<td>85.01</td>
<td>82.63</td>
<td>83.52</td>
<td>71.36</td>
<td>81.06</td>
</tr>
<tr>
<td>Xiao &amp; Lee</td>
<td>83.21</td>
<td>88.13</td>
<td>82.71</td>
<td>72.76</td>
<td>93.68</td>
<td>88.26</td>
<td>86.16</td>
<td>86.46</td>
<td>90.23</td>
<td>75.05</td>
<td>84.66</td>
</tr>
<tr>
<td>Ours</td>
<td>83.94</td>
<td>92.58</td>
<td>90.23</td>
<td>71.21</td>
<td>96.55</td>
<td>91.28</td>
<td>84.75</td>
<td>89.85</td>
<td>87.89</td>
<td>80.81</td>
<td>86.91</td>
</tr>
</tbody>
</table>

Shoe-Attribute Open Pointy Sporty Avg

<table>
<thead>
<tr>
<th>LFW-10</th>
<th>Open</th>
<th>Pointy</th>
<th>Sporty</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>P &amp; G + CNN</td>
<td>77.1</td>
<td>72.5</td>
<td>71.6</td>
<td>73.7</td>
</tr>
<tr>
<td>Sandeep et al.</td>
<td>80.2</td>
<td>82.5</td>
<td>88.1</td>
<td>83.6</td>
</tr>
<tr>
<td>Ours</td>
<td>89.3</td>
<td>82.5</td>
<td>93.6</td>
<td>88.5</td>
</tr>
</tbody>
</table>

Ablation Study (LFW-10)

<table>
<thead>
<tr>
<th>LFW-10</th>
<th>BH</th>
<th>DH</th>
<th>EO</th>
<th>GL</th>
<th>ML</th>
<th>MO</th>
<th>S</th>
<th>VT</th>
<th>VF</th>
<th>Y</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Image</td>
<td>84.3</td>
<td>90.2</td>
<td>82.7</td>
<td>70.0</td>
<td>94.8</td>
<td>80.8</td>
<td>80.8</td>
<td>79.4</td>
<td>85.6</td>
<td>77.2</td>
<td>82.5</td>
</tr>
<tr>
<td>STN output</td>
<td>78.1</td>
<td>89.3</td>
<td>83.2</td>
<td>66.6</td>
<td>95.4</td>
<td>91.3</td>
<td>84.5</td>
<td>88.6</td>
<td>85.6</td>
<td>73.8</td>
<td>84.6</td>
</tr>
<tr>
<td>Combined</td>
<td>83.9</td>
<td>92.6</td>
<td>90.2</td>
<td>71.2</td>
<td>96.6</td>
<td>91.3</td>
<td>84.8</td>
<td>89.9</td>
<td>87.9</td>
<td>80.8</td>
<td>86.0</td>
</tr>
</tbody>
</table>

Acknowledgement: This work was supported in part by an Amazon Web Services Education Research Grant and GPUs donated by NVIDIA.
Efficient Large Scale Image Classification via Prediction Score Decomposition

Duy-Dinh Le¹,², Tien-Dung Mai², Shin’ichi Satoh¹, Thanh Duc Ngo², Duc Anh Duong²
(¹) National Institute of Informatics (NII), Japan
(²) University of Information Technology (UIT), Vietnam

Introduction

Task: Multi-class Classification
Given an image, predict which class that it belongs to.

Goal: Reducing the test time complexity necessary when the number of classes is large.

Challenges:
- The complexity in test time grows linearly with the number of classes when using the standard OvA approach.
- The label tree approach usually suffers the well-known error propagation problem and it is difficult for parallelization for further speedup.

Proposal:
- The key idea is to use a smaller number of classifier evaluations to reduce the testing cost.
- It is similar to the label tree approach, but hierarchical structure is not used.
- It is done through prediction score matrix decomposition.

Method Overview

Let’s revisit SVD.

\[ R = U \Sigma V^T \]

(2)

\[ R = U \Sigma V^T = U \Sigma U^T \]

(3)

Instead of obtaining \( U \) directly by singular value decomposition, we take into account that \( U \) is the result of performing regression on the feature vectors of the images. To do so, we will pose the original problem as an eigenvalue problem.

\[ R^T U = \lambda U \]

(4)

where \( U \) is an eigenvector and \( \lambda \) is the corresponding eigenvalue.

Now we consider an eigenvector setting, namely, \( u_i(x) = g(x) \). We further assume linear regression.

\[ p_i(x) = \langle u_i(x) \rangle = b_j + \langle u_j(x) \rangle \sum_{i=1}^{n} a_i \]

(7)

By defining the matrix of features

\[ S = [s_1, s_2, \ldots, s_n] \]

(9)

we want \( u = SS^T \). Substituting this into (6), the problem becomes:

\[ \min \| SS^T u - \lambda u \| \quad \text{s.t.} \quad u^T u = 1 \]

We can use the Lagrange multipliers method to solve it.

\[ \frac{\partial}{\partial u} (\| SS^T u - \lambda u \|^2) = 2SS^T u - 2\lambda u = 0 \]

\[ SS^T u = \lambda u \]

(11)

The above can be regarded as a constrained eigenvalue problem: \( P u = \lambda Q u \)

\[ P = SS^T \quad \text{and} \quad Q = S^T S \]

Training Stage

Obtain: \( P = SS^T, Q = S^T S \)

Solve the generalized eigenvalue problem: \( P u = \lambda Q u \)

Testing Stage

Verifying by \( Q \) OvA Classifiers

Effect of Quality of the Score Matrix \( R \) (training)

Conclusion

Propose an efficient method for large scale image classification.

The method is based on joint optimization for prediction score matrix decomposition.

Comprehensive evaluations on large datasets such as ImageNet-1K, ImageNet-10K show superiority over SoA methods such as OvA and label tree methods.
We developed a new algorithm for metric learning with covariance descriptors.

**Introduction**

The purpose of this study
- Improvement of generalization performance of the nearest neighbor classification for the covariance descriptor

Covariance Descriptor

Example: Video composed of T frames \((x_1, \ldots, x_T)\)

\[ X = \frac{1}{T} \sum_{t=1}^{T} (x_t - m)(x_t - m)^T \] Where, \( m = \frac{1}{T} \sum_{t=1}^{T} x_t \)

Pipeline for Nearest Neighbor Classification

1. Gather Labeled Data
2. Compute Distances to Labeled Data
3. Find the Nearest Neighbor (NN)

**Formalization of Learning Problem**

We propose the following learning problem.

**Objective Function**

\[ \min_{\mathbf{W}, \xi} \mathbb{B}D_2(\mathbf{W}, \xi), (I, 1) \]

**Constraints**

\( (\mathbf{W}, \xi) \in C_{\mathcal{L}} \)

- Half-spaces associated with class

\( \mathcal{C}_1 : = \{ (\mathbf{W}, \xi) | \mathbb{B}D_2(\mathbf{W}, \xi, I) \leq b_{\text{up}} \} \)

- Half-spaces associated with different class

\( \mathcal{C}_2 : = \{ (\mathbf{W}, \xi) | \mathbb{B}D_2(\mathbf{W}, \xi, W) \geq b_{\text{up}} \} \)

**Bregman Divergence**

\( \mathbb{B}D_2(\mathbf{W}, \xi, W) = \mathbb{B}D_2(\mathbf{W}, \xi) + \mathbb{B}D_2(W, W) \)

**Seed Function**

\( \omega(\mathbf{W}, \xi) := -\log(\mathbb{B}D_2(\mathbf{W}, \xi, W)) = \sum_{k=1}^{K} \omega_k(\xi_k) \)

**Dykstra Algorithm**

Projection onto the intersection can be found by repeatedly projecting onto each half-space.

Naive method takes \( O(Ln^2) \) to find the projection onto each \( \mathcal{C}_k \).

**Acceleration of Each Iterate**

Projection onto \( \mathcal{C}_k \)

\[ \min_{\mathbf{B}D_2(\mathbf{W}_k, \xi_k), (I-1)} \text{wrt } (\mathbf{W}_k, \xi_k) \in b(\mathcal{C}_k) \]

Where, \( \mathcal{C}_k : = \{ (\mathbf{W}, \xi) | \mathbb{B}D_2(\mathbf{W}, \xi, I) \leq \theta \} \)

Use Newton method to find a zero of the function \( f_1(\delta) = 0 \).

Naive

\( \delta(\beta) := \left( A_k^{-1} + \beta_k A_k \right)^{-1} \left( \beta_k \nabla^2 \left( \mathbb{B}D_2(\mathbf{W}, \xi, W) \right) + \beta_k \mathbf{W} \right) \)

Proposed

Compute the spectral decomposition of \( A_k = \mathbb{B}D_2(\mathbf{W}, \xi, I) \)

Preprocess

\( \delta(\beta) := \sum_{l=1}^{L} \frac{1}{\mathbf{g}_l} \beta_k \mathbf{W} \left( \nabla^2 \left( \mathbb{B}D_2(\mathbf{W}, \xi, W) \right) + \beta_k \mathbf{W} \right) \)

**Acceleration of Convergence**

**Proposal Methods - Stochastic Dykstra Algorithm**

- Conventional Dykstra Algorithm
  - When \( k = 4, \)

*Epoch1*  *Epoch2*  *Epoch3*

Proposed \( \mathcal{C}_1 \)  \( \mathcal{C}_2 \)  \( \mathcal{C}_1 \)

Proposed \( \mathcal{C}_2 \)  \( \mathcal{C}_1 \)  \( \mathcal{C}_2 \)

Proposed \( \mathcal{C}_2 \)  \( \mathcal{C}_1 \)  \( \mathcal{C}_2 \)

- Stochastic Dykstra Algorithm
  - When \( k = 4, \)

*Epoch1*  *Epoch2*  *Epoch3*

Proposed \( \mathcal{C}_1 \)  \( \mathcal{C}_2 \)  \( \mathcal{C}_1 \)

Proposed \( \mathcal{C}_2 \)  \( \mathcal{C}_1 \)  \( \mathcal{C}_2 \)

Proposed \( \mathcal{C}_2 \)  \( \mathcal{C}_1 \)  \( \mathcal{C}_2 \)

pattern Recognition Performance

- We did metric learning with four types of mapping functions.
- We used the following three objectives for learning.

**Mapping Functions**

- ISBD: \( \phi_1(\xi_k) := -\log(\xi_k) \)
- L2BD: \( \phi_2(\xi_k) := \frac{1}{2} \xi_k^2 \)
- REBD: \( \phi_3(\xi_k) := (\log \xi_k - 1) \xi_k \)

**Seed Functions**

- Texture recognition
- General object recognition

Generalization performance is improved by metric learning.
**MeshFlow: Minimum Latency Online Video Stabilization**

**Authors:**
- Shuaicheng Liu
- Ping Tan
- Lu Yuan
- Jian Sun
- Bing Zeng

**Institutions:**
1. University of Electronic Science and Technology of China
2. Simon Fraser University
3. Microsoft Research

**Video Stabilization Requirements:**
- Spatially-variant motion representation for high quality stabilization.
- Adaptive path smoothing for robust video stabilization.
- Online processing for real-time applications.

**Our Contributions:**
- PAPS: a Predicted Adaptive Path Smoothing method.
- MeshFlow + PAPS – minimum latency online video stabilization.

**MeshFlow motion model:**

1. **Vertex profile**
   - (a) A pixel profile in SteadyFlow
   - (b) A vertex profile in MeshFlow
   - (a) Pixel profiles collect motion vectors at the same pixel location in SteadyFlow over time for all pixel locations. Motions of SteadyFlow come from dense optical flow. (b) Vertex profiles only collect motion vectors in MeshFlow at mesh vertices. Motions of MeshFlow come from feature matches between adjacent frames.

2. **Model estimation**
   - A pair of matched features (red dots) between frame 1 and 2. The arrow indicates the motion of the feature point at frame 1. The motion is propagated to the nearby vertices.

3. **Predict \( \lambda_t \) for online stabilization**
   - The plot of \( \lambda_t \) and values of translational elements \( \mathbf{T} \) using 50 videos with quick camera motions (e.g., quick rotation, fast zooming).
   - (b) The plot of \( \lambda_t \) and values of affine component \( \mathbf{F} \) using another 50 videos containing large depth variations.

**Predicted Adaptive Path Smoothing (PAPS)**

1. **Offline stabilization**
   \[
   O(P(t)) = \sum_{r \in \Omega} \alpha_r (|P(t) - C(t)|^2 + \lambda_r (|P(t) - P(r)|^2))
   \]
   - \( C(t) \) The original camera path
   - \( P(t) \) Stabilized camera path
   - The adaptive weight \( \lambda \) is very important in video stabilization, which can adaptively control the smoothness, such that some artifacts can be suppressed effectively (e.g., large cropping, wobbling). It is a tradeoff between stability and some side-effects.

2. **\( \lambda_t \) calculation in offline stabilization**
   - The method in [2] adopted an iterative refinement approach to search the optimal value of \( \lambda_t \) for each frame. They proposed to evaluate the cropping and wobbling numerically. Although, this dynamic parameter adjustment can find the optimal values of \( \lambda_t \) but is not efficient. More importantly, it requires too many future frames.
   - Not efficient, require many future frames

**One frame delay online stabilization**

- The system is delayed at the time 0. At the time 1, we begin to display. The stabilization is started at the time 3 when there is at least one historical frame. The \( \xi \) indexes the optimization conducted at each time. The third term in Eq. 2 enforces similarities between paths in the current optimization \( \xi \) and paths obtained in the previous optimization \( 1 - \xi \), which is indicated by the dashed lines.

\[
\min_{P(t)} \left\{ \sum_{r \in \Omega} \alpha_r (|P(t) - C(t)|^2 + \lambda_r (|P(t) - P(r)|^2)) + \sum_{r \in \Omega} \alpha_r (|P(t) - C(t)|^2 + \lambda_r (|P(t) - P(r)|^2)) \right\}
\]

**Results**

Our experiments on various devices. Each column shows a capturing device and the corresponding sample frame. Please refer to the project page for the videos: [http://www.liushuaicheng.org/eccv2016/index.html](http://www.liushuaicheng.org/eccv2016/index.html)

**Flow comparison**

- (a) raw optical flow [3]
- (b) SteadyFlow [1]
- (c) Our interpolated mesh flow

Our MeshFlow is quite similar to the SteadyFlow. The MeshFlow enjoys the merits of the SteadyFlow while the computation is much cheaper.

---

For a given GMM, and a set of vectors \( X = (x_1, \ldots, x_n) \in \mathbb{R}^d \)

Consider the log-likelihood of \( X \) given the parameters of the model \( \lambda = (\mu, \Sigma) \):

\[
L(\lambda | X) = \sum_{i=1}^{n} \log(p(X_i | \lambda))
\]

For each parameter \( \ell \in \lambda \), we can compute:

\[
\frac{\partial L(\lambda | X)}{\partial \ell}
\]

This representation does not take the elements ordering into account.

To overcome this insensitivity to ordering, we replace the GMM by an RNN model.

**RNN Training**

We train an RNN to predict the next element in a sequence, given the previous ones.

Original sequence: \((x_1, \ldots, x_n)\) \((x_i \in \mathbb{R}^d)\)

Input sequence: \(X = (x_0, x_1, \ldots, x_{n-1})\)

Target sequence: \(Y = (x_1, x_2, \ldots, x_n)\)

Loss function: \(Loss(Y, \hat{y}) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2\)

**RNN Feature Extraction**

Given a new sequence \((x_1, \ldots, x_n)\), feed it to the RNN.

At time step \(t \), the output of the RNN is \(v_t\), and the likelihood of a vector \(x \in \mathbb{R}^d\) can be seen as:

\[
p(x|v_0, \ldots, v_{t-1}) = \frac{1}{Z} \exp\left(-\frac{1}{2} (x - v_t)^T \Sigma^{-1} (x - v_t)\right)
\]

The likelihood of the correct next word \(x_{t+1}\) is:

\[
p(x_{t+1}|v_0, \ldots, v_t) = \frac{1}{Z} \exp\left(-\frac{1}{2} (x_{t+1} - v_{t+1})^T \Sigma^{-1} (x_{t+1} - v_{t+1})\right)
\]

The likelihood of the entire sequence \(X\) is:

\[
p(X) = \prod_{t=1}^{N} p(x_t|x_0, \ldots, x_{t-1})
\]

The negative log-likelihood of \(X\):

\[
L(X) = -\frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{t=1}^{N} v_t^T \Sigma v_t - \frac{1}{2} \sum_{t=1}^{N} (x_t - v_t)^T \Sigma^{-1} (x_t - v_t)
\]

If the score \(L(X)\) is close to zero, the model could be improved by feeding \(X\) back to the RNN.

**Image-Sentence Retrieval Pipeline**

Sentence representation: RNN-FV

- RNN may be trained on external corpus (e.g., Wikipedia)
- Generic model


**Video Action Recognition**

**Image Search**

A man is walking down the street

An image is sensitive to order of elements

**Image-Sentence Retrieval**

**Video Action Recognition - Results**

Accuracy results on the HMDB51 and UCF101 datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>HMDB51</th>
<th>UCF101</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNN-FV vs. GMM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG</td>
<td>66.78</td>
<td>29.9</td>
</tr>
<tr>
<td>VGG CCA</td>
<td>85.2</td>
<td>35.6</td>
</tr>
<tr>
<td>VGG PCA</td>
<td>70.0</td>
<td>30.0</td>
</tr>
<tr>
<td>SDT</td>
<td>79.0</td>
<td>35.6</td>
</tr>
<tr>
<td>NIC</td>
<td>66.7</td>
<td>29.9</td>
</tr>
<tr>
<td>map</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG CCA</td>
<td>56.8</td>
<td>22.4</td>
</tr>
<tr>
<td>VGG PCA</td>
<td>48.8</td>
<td>20.7</td>
</tr>
<tr>
<td>SDT</td>
<td>40.0</td>
<td>23.9</td>
</tr>
<tr>
<td>NIC</td>
<td>38.0</td>
<td>20.7</td>
</tr>
<tr>
<td>SDT</td>
<td>40.0</td>
<td>23.9</td>
</tr>
<tr>
<td>VGG FV (Ens)</td>
<td>56.8</td>
<td>22.4</td>
</tr>
<tr>
<td>VGG FV (Wikipedia-trained)</td>
<td>48.8</td>
<td>20.7</td>
</tr>
<tr>
<td>SDT</td>
<td>32.0</td>
<td>20.7</td>
</tr>
<tr>
<td>NIC</td>
<td>38.0</td>
<td>20.7</td>
</tr>
<tr>
<td>SDT</td>
<td>40.0</td>
<td>23.9</td>
</tr>
</tbody>
</table>

**Image-Sentence Retrieval - Results**

Recall@1 results on the flickr8k and flickr30k datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>Image</th>
<th>Image</th>
<th>Image</th>
<th>Image</th>
<th>Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG PCA</td>
<td>35.6</td>
<td>21.3</td>
<td>26.0</td>
<td>22.4</td>
<td></td>
</tr>
<tr>
<td>VGG CCA</td>
<td>56.8</td>
<td>22.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG PCA</td>
<td>48.8</td>
<td>20.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDT</td>
<td>32.0</td>
<td>20.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIC</td>
<td>38.0</td>
<td>20.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SDT</td>
<td>40.0</td>
<td>23.9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**References**

Motivation

- In tracking-by-detection, the tracking performance is closely related to the detection quality.
- A model-free tracking algorithm can be used to find undetected objects.

Our approach

- We improve the detection quality by combining an object detector with a model-free tracker.

Object Detection

Training

- We employ the pre-trained Fast R-CNN detector
- Then, we fine-tune the detector on MOT challenge dataset.

Test

- We use only the detection results with high confidence to reduce false positives.
- In the proposed algorithm, the impacts of undetected objects are less severe than those of false positives.

Experimental Results

- MOT challenge 2015 dataset

Analysis

- Detection performance on validation sequences
- Tracking performance on validation sequences

Comparison Results

- We use the same detector for comparison

Post-processing

- In the proposed algorithm, the impacts of undetected objects are less severe than those of false positives.

Tracking results

Overview

- We model the appearance of an active target using SSVM
- We perform model-free tracking to find target states
- Detection guidance
  - Find the disappearing target by computing the detection score
  - Determine occluded targets by computing IoU overlap ratio
- Matching between detection results and active targets
  - Improve target states using matched detection results
- Add new active target using unmatched detection results
- Appearance Model Update
  - We model the appearance of an active target using SSVM
MARS: A Video Benchmark for Video-based Person Re-identification

*Liang Zheng1,2, *Zhi Bie1, *Yifan Sun1, Jingdong Wang3, Shengjin Wang1, Chi Su1, Qi Tian2
1Tsinghua University 2University of Texas at San Antonio 3Microsoft Research 4Peking University * equal contribution

Introduction

Video Person Re-identification

Contribution

Current video re-id datasets typically
- have a very small scale.
- use hand-drawn bboxes for each video sequence.
- do not include false detection results on the background.
- have only one query for each identity.
- have only one ground truth for each identity.
- only use CMC curve for evaluation.

We introduce the MARS dataset
- is large scale: 1,261 IDs and 20,715 tracklets.
- uses DPM detector + GMMCP (CVPR15) tracker.
- has multiple queries for each identity.
- has multiple ground truths for each identity.
- uses mean Average Precision (mAP) & CMC.

We use the ID-discriminative Embedding (IDE)
- easy to train/test
- produces competitive accuracy on iLIDS-VID and new state of the art on PRID-2011.

MARS (Motion Analysis and Re-identification Set)

Comparison with current re-id datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>MARS</th>
<th>iLIDS</th>
<th>PRID</th>
<th>3DPEs</th>
<th>ETH</th>
<th>CUHK03</th>
<th>VIPER</th>
<th>Market-1501</th>
<th>RAID</th>
</tr>
</thead>
<tbody>
<tr>
<td>#identities</td>
<td>1,261</td>
<td>300</td>
<td>200</td>
<td>200</td>
<td>146</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>#tracklets</td>
<td>20,715</td>
<td>600</td>
<td>400</td>
<td>1,000</td>
<td>146</td>
<td>1,360</td>
<td>632</td>
<td>1,501</td>
<td>43</td>
</tr>
<tr>
<td>#bboxes</td>
<td>1,067,516</td>
<td>43,800</td>
<td>40k</td>
<td>200k</td>
<td>5,850</td>
<td>13,164</td>
<td>1,264</td>
<td>32k</td>
<td>6,920</td>
</tr>
<tr>
<td>#distractors</td>
<td>3,248</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2,793</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>#cam./ID</td>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>produced by</td>
<td>DPM+GMMCP</td>
<td>Hand</td>
<td>Hand</td>
<td>Hand</td>
<td>Hand</td>
<td>DPM</td>
<td>Hand</td>
<td>DPM</td>
<td>Hand</td>
</tr>
<tr>
<td>Evaluation</td>
<td>mAP+CMC</td>
<td>CMC</td>
<td>CMC</td>
<td>CMC</td>
<td>CMC</td>
<td>CMC</td>
<td>CMC</td>
<td>mAP</td>
<td>CMC</td>
</tr>
</tbody>
</table>

Sample sequences in MARS

Evaluation Details

✓ Cross-camera search.
✓ Fixed train/test partition
✓ Train: 625 IDs, 8,298 tracklets
✓ Gallery: 636 IDs, 12,180 tracklets.
✓ 2,007 queries, 3.7 ground truths per query.
✓ Given a query, we search in 12,180 tracklets, i.e., the test set

ID-discriminative Embedding (IDE)

✓ We train an AlexNet on the training set
✓ We classify each training bbox into one of the 625 IDs
✓ We extract FC7 for each bbox during testing
✓ We use max/avg pooling to obtain video features.

MARS statistics

Comparison with state of the art

Sample re-id results on MARS. (Row 1: motion feature, Row 2: IDE feature)

Method Comparison

Video re-id naturally outperforms image-based re-id

Motion features are less effective on MARS
Newtonian mechanics inspired methods such as Social Force Model have been successfully applied for anomaly detection in crowd scenes. However, several recent socio-psychology studies have shown that current SFM-based methods may not be capable of explaining behaviors in complex crowd scenarios. Thus, an alternative approach consists in describing the cognitive processes giving rise to the behavioral patterns observed in crowd using heuristics. Inspired by these studies, we propose a new hybrid framework to detect violent events in crowd scenarios.

H1: Acceleration
\[ \frac{dv}{dt} \approx a_{heading}^{t+1} - a_{heading}^{t} \]
\[ \frac{d^2v}{dt^2} : \text{Acceleration of Ped, at the time } t \]
\[ \text{Heading}: \text{Velocity of Ped, at the time } t \]

H2: Body Compression Force
\[ F^c = \sum_i \left( q_i - q_i^c \right) \cdot \left( r_{ij}^c \cdot c_i \right) \]
\[ q_i : \text{Contact force between Ped}_{i} \text{ and Ped}_{j} \]
\[ r_{ij}^c : \text{2D Gaussian function over acceleration map} \]

H3: Aggression Force
\[ F^{agg} = \frac{1}{2} \left( 1 \text{ - } \cos \phi_{ij} \right) \cdot \left( q_i - q_i^c \right) \cdot \left( r_{ij}^c \cdot c_i \right) \]
\[ Q : \text{Quantized bins of optical flow orientations} \]
\[ w^f_{ij} : \text{Aggression factor} \]

Two approximations used for aggression factors
\[ w^f_{ij} = \begin{cases} 1 & \text{if } \phi_{ij} = \phi_{ij}^c \\ 0 & \text{otherwise} \end{cases} \]
\[ w^h_{ij} = \begin{cases} 1 & \text{if } \phi_{ij} \neq \phi_{ij}^c \\ 0 & \text{otherwise} \end{cases} \]

1. Three behavioral heuristics are proposed.
2. Each heuristic is first formulated, and estimated from video.
3. Bag-of-Words paradigm is used for the representation of the estimated heuristics.
4. All the forces are concatenated to shape the final descriptor, named, Visual Information Processing Signature (VIPS).

**REFERENCES**

Sparse Recovery of Hyperspectral Signal from Natural RGB Images
Boaz Arad and Ohad Ben-Shahar
Department of Computer Science, Ben-Gurion University of the Negev

1. Hyperspectral Imaging
Hyperspectral imaging systems (HIS) collect scene radiance over a wide range of narrow bands.

Airborne HS
NASA AAT (9)

Laboratory HS
Nuance FX imaging system

Traditional HIS employ either spatial or spectral scanning, are costly and unsuited for imaging of natural scenes.

2. Related Work
A variety of systems have attempted to overcome the limitations of traditional HIS

Computed Tomography
Hybrid RGB/HS
Hybrid Compressive RGB/HS

3. Hyperspectral from RGB
Could we replace slow, expensive HIS with fast, low cost consumer cameras?

~31 dimensional signal from
3 dimensional input!

Hyperspectral signals in nature are sparse both spatially and spectrally (6).

Leveraging the sparsity of hyperspectral images, the task of Hyperspectral-from-RGB becomes possible.

4. Natural Hyperspectral Database
Due to the difficulty involved in acquiring HS images of natural scenes, only a few such databases exist.

Notable hyperspectral databases:

<table>
<thead>
<tr>
<th>Database</th>
<th>Images</th>
<th>Spatial Resolution</th>
<th>Spectral Resolution</th>
<th>Image Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>33</td>
<td>100x100</td>
<td>100x100</td>
<td>Urban, Rural</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>256x256</td>
<td>256x256</td>
<td>Rural, Farm</td>
</tr>
<tr>
<td>B</td>
<td>12</td>
<td>1024x1024</td>
<td>1024x1024</td>
<td>Urban, Industrial</td>
</tr>
<tr>
<td>C</td>
<td>20</td>
<td>1340x1340</td>
<td>1340x1340</td>
<td>Studio</td>
</tr>
<tr>
<td>D</td>
<td>30</td>
<td>1392x1300</td>
<td>1392x1300</td>
<td>Highly Varied</td>
</tr>
</tbody>
</table>

5. Approach
A Hyperspectral prior is reduced to a sparse, overcomplete dictionary $D_h$:

$D_h = \{h_1, h_2, ..., h_n\}$

This dictionary is then projected to RGB via the target cameras known response function ($R$):

$D_{rgb} = \{c_1, c_2, ..., c_n\} = R \cdot D_h$

Given an RGB query signal $c_q$, find $w$ such that:

$D_{rgb} \cdot w = c_q$

Once $w$ is found, the hyperspectral signal $c_h$ is estimated:

$c_h = D_h \cdot w$

In our implementation, $D_h$ is created via K-SVD (13) and is found via OMP (14).
For simulated camera experiments, the CIE color matching function is selected.

6. Results - Simulated Camera
Our methodology achieves >90% accuracy over never-before-seen images in simulated test. Results are on-par with those of some hybrid HS-RGB systems (12).

Average relative RMSE: 7.2%

7. Results - Simulated Camera
Similarly accurate results were achieved using RGB input from a consumer camera in both real-world and simulated tests.

Average relative RMSE: 7.5%

8. Summary
By accurately estimating hyperspectral signals from RGB images, we present a low cost, instantaneous, single shot alternative to scanning HS imaging systems.

Our reconstruction method is computationally efficient (seconds on single core) and allows video-rate HS-from-RGB via parallelization and/or memoization.

A larger-than-ever natural hyperspectral image database provides a rich prior for our reconstruction method and facilitates future research.

Bibliography

Acknowledgments

Scan code for database
Light Field Segmentation Using a Ray-Based Graph Structure

Matthieu Hog\textsuperscript{1,2}, Neus Sabater\textsuperscript{1}, Christine Guillemot\textsuperscript{2}
\textsuperscript{1}Technicolor R\&I, \textsuperscript{2}INRIA , Rennes, France

Motivation and Problem

- We are interested in interactive light field segmentation with a reference view.
- Challenge for light fields: the amount of data to process in order to edit all the views simultaneously.
- In particular, the running time for graph-based approaches, such as graph-cut, increases greatly with the size of the input graph.
- We assume a depth map to be known for each view.

Our Method

Building The Graph

- Previous work focuses on representations with one graph node per ray [1].
- Observation: many rays of the light field mainly describe the same content.
- The redundancy is captured by depth estimation.
- We use a single node to represent several rays coming from the same scene point (ray bundles), according to an estimated measure:

\[
|s_t + s_i | D(s_t, t, s_i, t_i, y_j)|| = s_t
\]

- To handle occlusions and errors in the depth map, rays that have an incoherent depth measure (free rays) are left in a single node.
- The new neighbourhood relationship are defined using each view neighborhood.

New Energy Terms

- Ray bundle unary energy term is defined using Gaussian Mixture Models (GMM) on the color and the depth learnt from input scribbles. To compute the free ray unary, the depth component is removed from the GMM:

\[
U(b_i) = \begin{cases} \log \frac{P(C(b_i)) D(b_i)}{P(C(b_i)) D(b_i)} & \text{if } \exists r_i \in b_i, S(r_i) = 0 \\ \infty & \text{if } \exists r_i \in b_i, S(r_i) = \alpha \\ 0 & \text{otherwise} \end{cases}
\]

- We define the new neighbourhood relationship using the neighbourhood on each view. Between free rays and ray bundles:

\[
P(r_i/b_i, r_j) = \delta_{r_i/b_i} \cap r_j \exp \frac{-\Delta E(C(r_i/b_i), C(r_j))}{\delta_{lab}}
\]

- Problem: rays at the object boundaries tend to be more connected to the background node they occlude.
- Solution: sum the individual view neighbourhood energy term:

\[
P(b_i, b_j) = \delta_{b_i/b_j} \cap b_i \cap b_j \exp \left( \frac{-\Delta E(C(b_i), C(b_j))}{\delta_{lab}} \right) \left( \frac{D(b_i) - D(b_j)}{\delta_D} \right)
\]

- We use alpha-expansion to minimise the new energy function:

\[
\phi_{L}(b_i) = \sum_{r_i \in b_i} U(r_i) + \sum_{b_i \in B} U(b_i) + \sum_{r_i \in K, r_j \in K(b_i)} P(r_i, r_j) + \sum_{b_i \in B(b_i), b_j \in B(b_j) \cap b_i} P(b_i, b_j)
\]

Experimental Results

- Quantitative tests on synthetic dense light field dataset showed our approach to be close to the state of the art [1], with a lower complexity. Our approach also shows to be efficient for dense and sparse real light fields from various sources.

Summary and Conclusions

- We proposed a new graph structure that greatly reduces the running time of graph-based algorithm for light fields.
- We demonstrate its efficiency for interactive light field segmentation.
- Limitations: as the amount of depth estimation errors increases, the number of nodes rises and the segmentation coherence decreases.

References

[1] Wanner et al., Globally consistent multi-label assignment on the ray space of 4d light fields, CVPR 13
Design of Kernels in Convolutional Neural Networks for Image Classification

Zhun Sun  Mete Ozay  Takayuki Okatani
Graduate School of Information Sciences, Tohoku University
sun, mozay, okatani@vision.is.tohoku.ac.jp

Abstract
Despite the effectiveness of convolutional neural networks (CNNs) for image classification, our understanding of the effect of shape of convolution kernels on learned representations is limited. In this work, we explore and employ the relationship between shape of kernels which define receptive fields (RFs) in CNNs for learning of feature representations and image classification. For this purpose, we propose a shape visualization method for visualization of pixel-wise classification score maps of learned features. Motivated by our experimental results, and observations reported in the literature for modeling of visual systems, we propose a novel design of shape of kernels for learning of representations in CNNs. In the experimental results, the proposed models achieved an outstanding performance in the classification task, competing to a base CNN models that introduces more parameters and computational time. Additionally, we examined the region of interest (ROI) of different models in the classification task and analyzed the robustness of the proposed method to occluded images. Our results indicate the effectiveness of the proposed approach.

Introduction
Following the success of convolutional neural networks (CNNs) for large scale image classification, remarkable efforts have been made to deliver state-of-the-art performance on this task. Along with more complex and elaborate architectures, lots of techniques concerning parameter initialization, optimization and regularization have also been developed to achieve better performance. Despite the fact that various aspects of CNNs have been investigated, design of the convolution kernels, which can be one of the fundamental problems, has been rarely studied. Some studies examined how size of kernels affects performance [1], leading to a recent trend of stacking small kernels (e.g. 1 × 3) in deep layers of CNNs. However, analysis of the shapes of kernels is mostly left untouched. Although there seems to be no latitude in designing the shape of convolution kernels intuitively (especially 1 × 3 kernels), in this work, we suggest that designing the shapes of kernels is feasible and practical. Specifically, we propose a method to use an asymmetric shape, which simulates hexagonal lattices, for convolution kernels, and then deploy kernels with this shape in different orientations for different layers of CNNs.

Main Contributions
1. We propose a method to design convolution kernels in CNNs, which is inspired by hexagonal lattice structures employed for solving various problems of computer vision and image processing.
2. We examine classification performance of CNNs equipped with our kernels, and compare the results with state-of-the-art CNNs equipped with square kernels using benchmark datasets. The experimental results show that the proposed method is superior to the state-of-the-art CNN models in terms of computational time and/or classification performance.
3. We introduce a method for visualization of features to qualitatively analyze the effect of kernel design on classification. Additionally, we analyze the robustness of CNNs equipped with and without our kernel design to occlusion by measuring their classification accuracy when some regions on input images are occluded.

Our approach
In this work, we address the aforementioned problems by designing shapes of kernels on a two-dimensional coordinate system. For each channel of a given image I, we associate each pixel Ii,j ∈ I at each coordinate (i, j) with a lattice point (i.e., a point with integer coordinates) in a square grid. If two lattice points in the grid are distinct and each (i, j) differs from the coordinate point of the other by at most t, then they are called t-adjacent. An 8-neighbor of a lattice point Ii,j ∈ I is a point in 8-adjacent to Ii,j. We define N(Ii,j), as a set consisting of a pixel Ii,j ∈ I, and its 8 nearest neighbors (Fig. 1). A shape of a quasi-hexagonal kernel K(I(Dp,q)) ⊂ N(I[0,0]) is defined as

K(I(Dp,q)) = {N(Ii,j) | N(Ii,j) ∩ N(I[p,q]) ≠ ∅} (1)

where Dp,q ∈ D is a random variable used as an indicator function employed for designing of shape of K(I(Dp,q)), and takes values from D = {−1, 0, 1} | 0, 1, 0, 1}. Then, convolution of the proposed quasi-hexagonal kernel K(I(Dp,q)) on a neighborhood centered at a pixel located at (x, y) on an image I is defined as

I(x, y) ∗ K(I(Dp,q)) = ∑ ∑ I(i,j)K(I(Dp,q))(x-i,y-j) (2)

Results
We provide a comparison of the number of parameters and computational time of the models in Table 1. In the experimental analyzes for the CIFAR-10 dataset, QI-H model has a comparable performance to the base model with fewer parameters and computational time. If we keep the same number of parameters (QH-B), then classification accuracy improves for similar computational time. Meanwhile, our proposed model shows significant improvement in both model size and computational time in ImageNet dataset.

Table 1: Comparison of number of parameters and computational time of different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Num. of params.</th>
<th>Training time (500 samples)</th>
<th>Difference in accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE</td>
<td>≈ 57.3 M</td>
<td>≈ 64.1 M</td>
<td>95.3%</td>
</tr>
<tr>
<td>QI-H BASE</td>
<td>≈ 1.3 M</td>
<td>1277.4 ms</td>
<td>+1.8%</td>
</tr>
<tr>
<td>QI-H</td>
<td>≈ 1.3 M</td>
<td>1449.9 ms</td>
<td>+0.7%</td>
</tr>
</tbody>
</table>

References

Acknowledgements
This work was partly supported by CREST, JST and by JSPS KAKENHI Grant Number 15H01919 (Grant-in-Aid for Scientific Research on Innovative Areas Innovative SHITSUKSAN Science and Technology).
Learning Visual Features from Large Weakly Supervised Data
Armand Joulin*, Laurens van der Maaten*, Allan Jabri, Nicolas Vasilache
Facebook AI Research

1. Introduction
Learning good visual features requires increasingly large collections of manually labeled images. Manual labeling of images is cumbersome, error-prone and fails to exploit the massive amounts of weakly labeled data available on the internet. We train convolutional networks to predict words in user comments for a collection of 100 million Flickr images. We show the resulting networks learn visual features of high quality.

2. Dataset
The YFCC100M dataset contains 100 million images and associated user comments:

3. Models
- **Architectures**: AlexNet and GoogLeNet.
- **Loss function**: Cross-entropy loss (outperforms one-vs-all losses).
- **Training**: Stochastic gradient descent with batch size 128 (on 4 GPUs).
- **Class balancing**: Sample images uniformly per word.
- **Dictionary size**: Our models have up to 100,000 outputs.

To speed up learning, we use “stochastic gradient descent over words”. We bound the error of this approximation on both sides.

4. Associated Word Prediction
We measure word prediction quality (prec@10) on a test set of 1M images. Our baselines are Imagenet models with a linear layer for word prediction.

5. t-SNE Map
We visualize the visual features in a t-SNE map:

6. Transfer Learning
We transferred the visual feature to a range of other computer-vision tasks:

7. Assessing Word Embeddings
The output layer of our models contains word embeddings. We visualize the word embeddings in a t-SNE map:

8. Conclusion
Do we need to annotate millions of images to learn good vision models?

* Both authors contributed equally.
For more information, contact {ajoulin, lvdmaaten, ajabri, ntv}@fb.com.
### 3D Mask Face Anti-spoofing with Remote Photoplethysmography

Siqi Liu¹, Pong C. Yuen¹, Shengping Zhang², Guoying Zhao³

¹Department of Computer Science, Hong Kong Baptist University
²School of Computer Science and Technology, Harbin Institute of Technology
³Department of Computer Science and Engineering, University of Oulu

### MOTIVATION

**Background**
- Face spoofing attack is a critical security concern of the face recognition system.
- Existing methods perform well on traditional spoofing attack, e.g., prints attack / replay attack.
- 3D mask becomes the new challenge [1]

**3D Mask Challenge**
- Similar geometric property
- Similar appearance (color & texture)

**Problem Statement**
*How can we detect the 3D mask attack effectively?*

**MAIN IDEA**
- Using rPPG for 3D mask face anti-spoofing
- rPPG (remote Photoplethysmography)

- **Why does local rPPG work for 3D mask face anti-spoofing?**
  - rPPG on genuine face and masked face

- **Why use local rPPG?**
  - rPPG signal strength varies along local face region [2]
  - local rPPG strength forms a stable spatial pattern along different subjects ⇒ *discriminative structural information*

### PROPOSED METHOD

**Overview**
- Learning Local rPPG confidence map
- Transfer the confidence map into the distance metric in classifier

**Our new dataset:** A supplementary dataset to increase the varieties of 3DMAD
- 2 mask types (ThatsMyFace, REAL-F)
- 8 subjects
- recorded through web-cam

### EXPERIMENTS

**Effectiveness evaluation**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Combi</th>
<th>Suppl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Samples</td>
<td>Testing Samples</td>
<td>Training Samples</td>
</tr>
<tr>
<td>MS-LBP [3]</td>
<td>13.1±1.3</td>
<td>13.8±1.94</td>
</tr>
<tr>
<td>Proposed</td>
<td>9.2±2.0</td>
<td>9.7±12.6</td>
</tr>
</tbody>
</table>

**Robustness evaluation**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Combi</th>
<th>Suppl.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training/Testing Subjects</td>
<td>Training/Testing Samples</td>
<td>Training/Testing Subjects</td>
</tr>
<tr>
<td>MS-LBP [3]</td>
<td>30.6±5.7</td>
<td>49.2±6.9</td>
</tr>
<tr>
<td>Proposed</td>
<td>11.9±2.7</td>
<td>12.3±3.3</td>
</tr>
</tbody>
</table>

### References

Spoofing face recognition with 3D masks.

Robustness evaluation for remote photoplethysmography.
In Biometric (ICB), 2010 international conference on, pages 197–207.

Face spoofing detection from single images using infrared texture analysis.
In Biometrics (ICB), 2011 international conference on, pages 71–79.
IEEE, 2011.
Guided Matching based on Statistical Optical Flow
for Fast and Robust Correspondence Analysis

Josef Maier¹, Martin Humenberger¹, Markus Murschitz¹, Oliver Zende¹, and Markus Vincze²

¹ AIT Austrian Institute of Technology, (josef.maier/ml, martin.humenberger, markus.murschitz, oliver.zende ) @ait.ac.at
² Vienna University of Technology, vincze@acin.tuwien.ac.at

Abstract

Inspired by recent efforts in optimizing the matching process using geometric and statistical properties, our approach constrains the search space by utilizing spatial statistics from a small subset of matched and filtered correspondences. We call this method Guided Matching based on Statistical Optical Flow (GM-SOF). To ensure broad applicability, our approach works on high dimensional descriptors like SIFT but also on binary descriptors like FREAK. To evaluate our algorithm, we developed a novel method for determining ground truth matches, including true negatives, using spatial ground truth information of well known datasets. Therefore, we evaluate not only with precision and recall but also with accuracy and fall-out. We compare our approach in detail to several relevant state-of-the-art algorithms using these metrics. Experiments show that our method outperforms all tested solutions in terms of processing time while retaining a comparable level of matching quality.

Guided Matching based on Statistical Optical Flow

The goal of our algorithm is to significantly reduce the search space for feature matching.

1. We find the most distinctive features (shown as blue, green, and yellow crosses) in both images by local non-maxima suppression of their responses.
2. This allows fast similarity based matching on only a few features distributed over the whole image. The arrows in the figure represent the optical flow vectors resulting from these initial matches.

3. We calculate the Statistical Optical Flow (SOF) using the initial matches.
4. We use SOF (shown as arrows and dashed circles) to guide the matching of all remaining features, which are shown as red crosses.

This procedure leads to
- very fast matching compared to tested state-of-the-art methods while
- achieving similar matching quality

What is SOF?

The Statistical Optical Flow (SOF) is used to guide the matching process. Therefore, an initial search position is estimated in the second image and the search range is reduced to a small area.

It consists of statistics (displacement & uncertainty information) about the spatial displacements of the initial matches and is independently and sparsely estimated for several areas of the image.

The Evaluation Framework

- The framework generates ground truth matches out of spatial ground truth information that includes true negatives necessary for accuracy and fall-out.
- We use datasets KITTI flow and disparity [1] as well as all possible image pair combinations of “bark”, “boat”, “graffity”, and “wall” from Mikolajczyk et al. [2].

Results

- We compare our algorithm (GM) to selected state-of-the-art algorithms, namely hierarchical clustering tree (HC), priority search k-means tree (HK), linear matching (Li), Locality Sensitive Hashing (LSH), and randomized KD-tree (RA) from the FLANN library [6], as well as CasHash (CH) [4] and SparseVFC (VFC) [5] in combination with the hierarchical clustering tree.
- Evaluations were performed in terms of processing time...

- and quality:

This table shows the evaluation results for the different algorithms:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time (ms)</th>
<th>Accuracy</th>
<th>Fall-out</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM</td>
<td>0.01</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>CH</td>
<td>0.1</td>
<td>0.98</td>
<td>0.02</td>
</tr>
<tr>
<td>HC</td>
<td>0.05</td>
<td>0.98</td>
<td>0.03</td>
</tr>
<tr>
<td>HK</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
</tr>
<tr>
<td>RA</td>
<td>0.03</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>LSH</td>
<td>0.04</td>
<td>0.98</td>
<td>0.02</td>
</tr>
<tr>
<td>VFC</td>
<td>0.1</td>
<td>0.97</td>
<td>0.02</td>
</tr>
</tbody>
</table>

We present - for the first time - accuracy ACC = ( TP + TN ) / ( P + N ) and fall-out FPR = FP / ( FP + TN ) values for all compared algorithms in addition to the usual measures precision and recall.

These measures are important, as accuracy enables to quantify the closeness of a matching algorithm’s output to the true solution, while fall-out is a direct measure on the algorithm’s failure rate in correlation with non-matchable keypoints (true negatives).

REFERENCES

The problem ...

High Level of Heterogeneity in the Object Detection & Pose Estimation Problem!

- Different Datasets:

- Different Evaluation Metrics:
  - Pose estimation & detection → separated evaluation.
  - Only in the last years, 3 different metrics for simultaneous evaluation.

Our proposal

1. To consolidate the work.
2. A publicly available diagnostic tool.
3. For the PASCAL 3D+ dataset:
   - Evaluation of the main false positives.
   - Influence of the object characteristics.
   - Precise evaluation of the metrics.

Diagnostic Tool

Our analysis

1. For pose estimation only, models are biased towards the training data distribution.

2. For simultaneous detection and pose estimation, the main difficulty is: to obtain a precise BB and a correct estimation for the frontal/rear views.

3. Size and Aspect Ratio matter: all methods present difficulties working with unusual aspect ratios and sizes of the objects.

4. There is a correlation between easy-to-detect objects and easy-to-estimate-pose objects.

5. Evaluation metrics.
   - AOS is greatly dominated by the detection performance.
   - AVP and PEAP are more adequate to simultaneously evaluate the detection and pose estimation performance.

Examples of how by using our diagnostic tool the models can be improved

Part-based Approaches

Models which decouple detection & pose estimation tasks

This work is supported by projects DGT SPIP2015-01809, MINECO TEC2013-45183-R, FWO G069612N (Representations and algorithms for the captation, visualization and manipulation of moving 3D objects), Nissan (1188371-1-UDARQ).
**Integration of Probabilistic Pose Estimates From Multiple Views**

ozgur.erkent, dadhichi.shukla, justus.piater @uibk.ac.at

UNIVERSITY OF INNSBRUCK, INTELLIGENT AND INTERACTIVE SYSTEMS GROUP

---

**Introduction**

- Problems in pose estimation:
  - Ambiguity in estimated pose of the object
  - The presence of outliers: similarity in the appearance from the observed view
- We introduce an approach that uses RGB-D images from different viewpoints
- Possible problems of multi-view integration:
  - Occlusions or unseen features in one of the views
  - Incomplete or noisy sensor information

**Probabilistic Appearance Based Estimation (PAPE)**

**Probabilistic Model of Appearance**

- Set of image features as a distribution [1]:
  \[
  \phi_i^e(x_i) = \int \mathcal{N}(p_{x_i}, p_{y_i}, a_i) K^v(a_i, a_{xy}) \, dy
  \]  

(1)

- Similarity of two distributions [1]:
  \[
  \left( \phi_i^e \ast \phi_j^e \right)(x_i) = \int \int \phi_i^e(x_i + y) \phi_j^e(y) \, dy \, dx
  \]  

(2)

- Combination of different features (absence of a feature with probability \(\lambda_i\)):
  \[
  \Phi_{x_i,v} = \prod_{f} \Phi_{x_i,v}^f (1 - \lambda_i^f) + \lambda_i^f
  \]  

(3)

**Integration of Feature Types**

**Multi-view Integration**

- All possible pose estimation combinations from view subsets: \(C = \{(x_{\alpha1}, \ldots, x_{\alphaN}) : x_{\alpha1} \in v_i, \forall v_i \in \mathcal{V}, \forall \alpha_i \in 1, \ldots, N^v_i\}\)
- Distribution of the combined pose estimates \(\varphi(c_j) = \prod v_i \in c_j \Phi(x_{\alpha1})\), where \(\Phi(x_{\alpha1}) = \mathcal{N}(x_{\alpha1}, \Sigma)\)
- Find the maximum score among combinations \(C: x^* = \arg \max_{x_{\alpha1}} \int \varphi(c_j) \lambda_{\alpha_i}\)

**Experiments**

**Single-View Pose Estimation**

**Multi-View Detection**

<table>
<thead>
<tr>
<th>AP in %</th>
<th>mDPM + mVFH [4]</th>
<th>mPAPE 3 Cams</th>
<th>mPAPE 1 Cam</th>
</tr>
</thead>
<tbody>
<tr>
<td>avocado</td>
<td>100.0</td>
<td>100.0</td>
<td>79.7</td>
</tr>
<tr>
<td>bowl</td>
<td>87.0</td>
<td>99.7</td>
<td>90.3</td>
</tr>
<tr>
<td>coffee box</td>
<td>80.0</td>
<td>92.4</td>
<td>80.0</td>
</tr>
<tr>
<td>coffee can</td>
<td>89.6</td>
<td>89.9</td>
<td>98.1</td>
</tr>
<tr>
<td>cup</td>
<td>100.0</td>
<td>97.6</td>
<td>82.8</td>
</tr>
<tr>
<td>nutella can</td>
<td>89.2</td>
<td>93.6</td>
<td>58.4</td>
</tr>
<tr>
<td>plate</td>
<td>90.2</td>
<td>98.1</td>
<td>82.3</td>
</tr>
<tr>
<td>spice can</td>
<td>98.5</td>
<td>96.8</td>
<td>78.0</td>
</tr>
<tr>
<td>sponge</td>
<td>97.0</td>
<td>97.4</td>
<td>48.3</td>
</tr>
<tr>
<td>mean</td>
<td>92.4</td>
<td>96.2</td>
<td>77.5</td>
</tr>
</tbody>
</table>

**Avg**

F1 Scores

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Joystick</td>
<td>51.5</td>
<td>53.4</td>
</tr>
<tr>
<td>Camera</td>
<td>80.7</td>
<td>37.2</td>
</tr>
<tr>
<td>Coffee Cup</td>
<td>99.5</td>
<td>87.7</td>
</tr>
<tr>
<td>Shampoo</td>
<td>82.5</td>
<td>75.9</td>
</tr>
<tr>
<td>Milk Carton</td>
<td>27.2</td>
<td>38.5</td>
</tr>
<tr>
<td>Juice Carton</td>
<td>41.2</td>
<td>87.0</td>
</tr>
<tr>
<td>Avg</td>
<td>63.8</td>
<td>63.3</td>
</tr>
</tbody>
</table>

**Contributions**

- Integration of pose estimates in 6DoF from multiple views even in the absence of a correct estimation in some of the views.
- Combination of different appearance-based attributes in the presence of noisy or incomplete data.

**References**


**Acknowledgements**

The research leading to these results has received funding from the European Community’s Seventh Framework Programme FP7/2007-2013 under grant agreement no. 610878, 3rd HAND.
Problem and Motivation

Ventricles from MRI & CT
Free 1-dim faces

Fast Marching produces:
• Fast Marching (Sethian 1996) solves eikonal equation:
  \[ \nabla U(x) = \phi(x) \quad x \text{ not } p \quad \phi(x) = \text{edge map} \]
  \[ U(p) = 0 \quad p = \text{seed point} \quad U(x) = \text{distance map} \]

• Fast Marching produces:
  - weighted distance \( U \) to seed point
  - fronts localized to curve
  - minimal paths from any point to seed point (by following gradient of \( U \))

• Local maxima of Euclidean length of minimal paths on front lies on curve of interest (Kaul & Yezzi 2011)

Intuition from 2D: Extracting a Curve

Fast Marching produces fronts (green)

Visual comparisons to state-of-the-art

Quantitative evaluation

Analysis under smoothing degradation

Robustness to seed point choice

Ridge Extraction

Algorithm (Deformation retraction of Fast Marching front to obtain ridge)
1. Form 2D cubical complex of front
2. Deform retract front to form ridge:
   a) Remove free faces in order based on Euclidean minimal path length (Euclidean path length easy to compute with Fast Marching)
   b) Stop when no free faces remain

Cubical d-complex -- set of faces of (of dim \( d \)) such that all sub-faces of each face are in the set Free face (from cubical complex theory):
• a face with a sub-face \( g \) such that \( g \) is not a sub-face of any other face
• intuition: a face that can be removed while preserving homotopy equivalence

Illustration of deformation retraction by ordered removal to obtain ridge

Surface Extraction

Algorithm (Deformation retraction of image to form surface)
1. Construct 3D cubicle complex of image
2. Remove 3D free faces in order of \( U \) distance from Fast marching
3. Remove 2D free faces not connected to given 1D cubicle complex of boundary curve in order of \( U \)
4. Until no free faces remain

Illustration of deformation retraction by ordered removal to form surface

Sample Results of SurfCut in Some Applications
Faults in Seismic Data Ventricles from MRI & CT Knee Cartilage from MRI Lung Fissures from CT

SurfCut: Free-Boundary Surface Extraction
Marei Algarni and Ganesh Sundaramoorthi
King Abdullah University of Science and Technology (KAUST)
CATS: Co-saliency Activated Tracklet Selection for Video Co-localization

Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan

Introduction

Goal: To localize the common object from set of similar videos, which is also known as video co-localization.

Challenges:
- Inter video variation:
- Intra video variation:

This Paper: Single video processing based on co-saliency object priors at regular intervals. These object priors are derived from inter video commonness, intra video commonness, and total motion saliency.

The Idea: Leverage co-saliency activated tracklets to perform video co-localization.

Proposed Method

1) Co-saliency Generation: Appropriate neighboring warped saliency maps are fused with the saliency maps of activators to generate different co-saliency maps. These maps are then similarly fused through averaging for generating eventual co-saliency object prior (O).
2) Bounding-Box Filtering: Co-saliency object prior helps in filtering out noisy bounding box proposals and keep good proposals.
3) Tracklets Generation: Tracklets are generated from good proposals up to the next activator, and are then scored using object priors at the two ends.
4) Tube Generation: Tracklets with high confidence scores forming good spatio-temporal consistency are chosen for tube generation.

Experimental Results

Results Considering Inter Video Variations

Final Co-saliency Maps

Results Considering Intra Video Variations

<table>
<thead>
<tr>
<th>WITH LABELS</th>
<th>Aircraft</th>
<th>Bird</th>
<th>Boat</th>
<th>Car</th>
<th>Cat</th>
<th>Cow</th>
<th>Dog</th>
<th>Horse</th>
<th>Motorcycle</th>
<th>Train</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg</td>
<td>81.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
<td>78.6</td>
</tr>
<tr>
<td>ext</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
<td>56.5</td>
</tr>
<tr>
<td>ext-vel-rot</td>
<td>60.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
</tr>
<tr>
<td>W/O LABELS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg</td>
<td>65.4</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
<td>61.8</td>
</tr>
<tr>
<td>ext</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
<td>55.2</td>
</tr>
<tr>
<td>ext-vel-rot</td>
<td>60.7</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
<td>57.4</td>
</tr>
</tbody>
</table>
Online Human Action Detection using Joint Classification-Regression Recurrent Neural Networks
Yanghao Li1, Cuiling Lan2, Junliang Xing3, Wenjun Zeng2, Chunfeng Yuan3, and Jiaying Liu1

1 Peking University  2 Microsoft Research Asia  3 Institute of Automation, Chinese Academy of Sciences
{llytonhao, liujiaying}@pku.edu.cn, {culan, wezeng}@microsoft.com, {jlxing, cfyuan}@nlpr.ia.ac.cn

### Problem and Framework

- **Task**: Online Action Detection
  - Detect the action type and location on the fly
  - Forecast the start and end points
- **Solution**: End-to-end Joint Classification-Regression RNN
  - Leverage long-range temporal dynamics
  - Frame-wise detection and forecasting
  - Avoid sliding-window design
  - A new action dataset\(^1\) for online action detection

### Joint Classification-Regression RNN

- **Deep LSTM Network**
  - FC Layer
  - LSTM Layer
  - FC Layer
  - LSTM Layer
  - FC Layer
  - LSTM Layer
- **Classification Task**
  - FC1
  - SoftMax
  - Action Label
- **Regression Task**
  - FC2
  - Soft Selector
  - FC3
  - Start/End Confidence

- Feature extraction and temporal dynamic modeling
- Frame-wise action classification
- Accurate localization & Start/End point forecasting

### Experimental Results

#### Action Detection

<table>
<thead>
<tr>
<th>Actions</th>
<th>SVM-SW</th>
<th>RNN-SW</th>
<th>CA</th>
<th>JCR-SW</th>
</tr>
</thead>
<tbody>
<tr>
<td>drinking</td>
<td>0.146</td>
<td>0.441</td>
<td>0.584</td>
<td>0.574</td>
</tr>
<tr>
<td>eating</td>
<td>0.465</td>
<td>0.550</td>
<td>0.588</td>
<td>0.533</td>
</tr>
<tr>
<td>writing</td>
<td>0.661</td>
<td>0.809</td>
<td>0.489</td>
<td>0.482</td>
</tr>
<tr>
<td>opening cupboard</td>
<td>0.396</td>
<td>0.321</td>
<td>0.460</td>
<td>0.495</td>
</tr>
<tr>
<td>washing hands</td>
<td>0.562</td>
<td>0.666</td>
<td>0.672</td>
<td>0.718</td>
</tr>
<tr>
<td>opening microwave</td>
<td>0.697</td>
<td>0.665</td>
<td>0.608</td>
<td>0.703</td>
</tr>
<tr>
<td>grilling</td>
<td>0.461</td>
<td>0.510</td>
<td>0.597</td>
<td>0.643</td>
</tr>
<tr>
<td>throwing trash</td>
<td>0.554</td>
<td>0.674</td>
<td>0.630</td>
<td>0.659</td>
</tr>
<tr>
<td>washing</td>
<td>0.887</td>
<td>0.747</td>
<td>0.391</td>
<td>0.780</td>
</tr>
<tr>
<td>average</td>
<td>0.540</td>
<td>0.600</td>
<td>0.596</td>
<td>0.653</td>
</tr>
</tbody>
</table>

**Evaluation metrics**
- Detection: F1-Score
  - Correct: IOU > 0.6
  - \( F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \)
- Localization: SL-Score (EL-Score)
  - Gaussian-like confidences
  - Start or end point
  - \( c_i^e = e^{-\frac{(t_i - s_i)^2}{2\sigma^2}} \)

**Final objective function**

\[
L(V) = L(V) + \lambda L(V)' = \frac{1}{N} \sum_{t=0}^{T} \left( \sum_{j=0}^{N} z_{ij} \ln P(Y_{ij} | V_y,...,V_t) \right) + \lambda \left( l(c_i', p_i') + l(c_i, p_i) \right)
\]

\( N \) \( t \) \( j \) \( z \) \( Y \) \( V \) \( e \) \( s \) \( c \) \( p \) \( l \) \( \lambda \) \( i \) \( j \) \( k \) \( m \) \( n \) \( o \) \( q \) \( r \) \( s \) \( t \) \( u \) \( v \) \( w \) \( x \) \( y \) \( z \)

#### Action Forecasting

- Forecast performance
- JCR-RNN outperforms others
- Confusion Matrix (start forecast)
- Still improve space
  - e.g., Eating and drinking
  - Gargling and washing hand

**Action Forecasting**

<table>
<thead>
<tr>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start Forecast</td>
</tr>
<tr>
<td>End Forecast</td>
</tr>
</tbody>
</table>

**Evaluation metrics**
- Detection: F1-Score
  - Precison: \( \frac{P}{P+R} \)
  - Recall: \( \frac{R}{P+R} \)
  - \( F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \)

\( P \) \( R \) \( F \)


ECCV 2016
Amsterdam
Netherlands
SyB3R: A Realistic Synthetic Benchmark for 3D Reconstruction from Images

Andreas Ley, Ronny Hänsch, and Olaf Hellwich
Computer Vision & Remote Sensing, Technical University of Berlin, Germany

Framework
- Blender Scene
- Python Script
- Script Generator
- Cycles
- Ground Truth
- HDR Images

Details
- Scene properties, e.g. Surface Texture
- Camera Properties, e.g. Depth of Field

Datasets

Image Processing
- Rotation Motion Blur
- Radial Distortion
- Auto Exposure
- Sensor Noise
- Tonemapping

Evaluation
- SFM/MVS Pipeline

Contact & Code
Andreas Ley, andreas.ley@tu-berlin.de
Ronny Hänsch, r.haensch@tu-berlin.de
http://andreas-ley.com/projects/SyB3R

Modelled based on real measurements

Variance: \( \sigma_{\text{noise}}^2 = a I_{\text{noise}} + b \)

\( a \) and \( b \) are fitted to the noise observed in real test images.

Camera curves modelled based on real images.

Real Noise
Synth. Noise
Gaussian IID Noise
ISO100, 100% Texture
ISO100, 25% Texture
A Deep Learning-Based Approach To Progressive Vehicle Re-identification For Urban Surveillance

Contributions

- We propose the PROVID, a deep learning-based progressive vehicle Re-Id approach, which treats vehicle Re-Id as the coarse-to-fine search in the feature space, and the near-to-distant search in the real world surveillance.
- The deep convolutional neural network is adopted to learn appearance features as the coarse filter, and the deep siamese neural network is designed and trained for plate verification in the fine search.
- A comprehensive vehicle Re-Id dataset, VeRi, is built with plate and spatiotemporal labels, which also contains more vehicles from practical traffic scene than existing datasets.

Approach

- For the appearance-based coarse filtering, we adopt the fusion model of low-level hand-crafted features and high-level semantic attributes which are learned from deep convolutional neural networks (CNNs).
- For number plate-based accurate search, a Siamese neural network (SNN) is trained with large numbers of plate images for license plate verification instead of recognizing the characters of the license plate.
- At last, a spatiotemporal relation model, STR, is utilized to re-rank vehicles to further improve the final results of vehicle Re-Id.

Evaluation and Conclusions

VeRi: Vehicle Re-identification Dataset

- The VeRi Dataset is one of the largest vehicle Re-Id dataset built from urban surveillance videos with diverse vehicle attributes, sufficient license plates, and accurate spatiotemporal information.

Evaluation on VeRi

- The PROVID method achieves 9.28% improvements in mAP, 10.94% in HIT@1, and 5.3% in HIT@5 compared with the state-of-the-art appearance-based model.
- We also evaluate the speed of progressive method (157 ms/query), which reduces 87.84% time cost than the strategy without progressive fusion (1,292 ms/query).

Conclusions:

- We propose a deep learning-based progressive vehicle Re-Id approach, which is able to instantly discover, locate, and track the target vehicles in huge amount of urban surveillance videos.
Generative Visual Manipulation on the Natural Image Manifold

Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman* and Alexei A. Efros
UC Berkeley           *Adobe

1. Image Manipulation is Hard!

The Lack of “Safety Wheels”:
- any less-than-perfect edit immediately makes the image look unrealistic.
- classic visual manipulation paradigm does not prevent the user from “falling off” the manifold of natural images

2. Learning Natural Image Manifold

Generative Adversarial Network (GAN)
[Goodfellow et al. 14’] [Radford et al. 15’]

3. Manipulating the Latent Vector

Objective: $z^* = \arg\min_{z \in \mathcal{Z}} \frac{1}{2} \text{E}_{y \sim \text{data}}[\text{D}(G(z), y)] + \frac{\lambda_z \lambda_\beta}{2} \|z - z_0\|^2_2.

4. Edit Transfer

Motion $(u, v)$+ Color $(\Delta \text{A})$: estimate per-pixel geometric and color variation

5. Interactive Image Generation

User edits

Reference
Deep Cascaded Bi-Network for Face Hallucination
Shizhan Zhu¹, Sifei Liu¹,², Chen Change Loy¹,³, Xiaou Tang¹,³
¹Department of Information Engineering, The Chinese University of Hong Kong
²University of California, Merced
³Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

Motivation

<table>
<thead>
<tr>
<th>General Super-Resolution</th>
<th>Existing Face Approaches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only recovering, no synthesizing</td>
<td>Inconsistent visual quality</td>
</tr>
<tr>
<td>Cannot handle very low resolution input faces</td>
<td>Visually dissimilar</td>
</tr>
<tr>
<td>Did not exploit face priors</td>
<td>Exemplar based, very slow</td>
</tr>
<tr>
<td>Assumes correct alignment</td>
<td></td>
</tr>
</tbody>
</table>

Framework

The dense correspondence field prediction steps

\[ p_3 = p_2 + f_2(p_2, p_1) \]
\[ p_3 = p_2 + f_2(p_1, p_3) \]
\[ p_3 = p_2 + f_3(p_3, p_1) \]
\[ p_3 = p_2 + f_5(p_3, p_1) \]

The face hallucination steps

Face Hallucination for Real Surveillance Video Frames:

Handling Pose Variations:

Failure Case Analysis:

SR Input Resolution Lower Bound Analysis:

Codes are available NOW! Get it NOW!!!
https://github.com/zhusz/ECCV16-CBN
Motivation:

- Hyperspectral Imagery Denoising:
  - Hyperspectral images (HSIs) often suffer from noise corruption.
  - Correlation across spectrum and spatial similarity in HSIs are useful for denosing, but how to fully exploit them together is seldom studied.

- The Proposed Method:

  A cluster sparsity field based HSIs denoising (CSFHD) method:
  - The CSF prior jointly models the correlation across spectrum and spatial similarity of an HSI with the sparse representation model.
  - The state-of-the-art denoising results are demonstrated.

The Proposed Method:

A. Observation Model

Given the observation model \( F = X + N = \Phi Y + N \), where the HSI is sparsely represented on a spectrum dictionary, we have the likelihood as

\[
p(Y|X,\Lambda) \propto \|X - \Phi Y\|^2_2 \exp \left(-\|X\|^2_2 \right), \quad \text{with } \Sigma = \text{diag}(\Lambda),
\]

B. Cluster Sparsity Field (CSF) Prior

A Markov random field model based prior is proposed for the sparse representation of an HSI to explore its intra-cluster structures as FIG 1.

\[
p(Y) = \frac{1}{Z} \exp \left(-\sum_i E_{\text{int}}(Y_i) \right), \quad \text{with } E_{\text{int}}(Y_i) = \phi(Y_i) + \psi(Y_i).
\]

- Structured Sparsity Potential \( \phi(Y_i) \):
  - The correlation across spectrum is modelled as the structured sparsity
  - The spatial similarity is modeled as intra-cluster graph structure

- Graph Structure Potential \( \psi(Y_i) \):
  - The spatial similarity is modeled as intra-cluster graph structure

\[
\phi(Y_i) = \frac{1}{2} \| Y_i \|^2_2, \quad \psi(Y_i) = \frac{1}{2} \sum_j \| Y_j \|^2_2, \quad \text{with } \Gamma_i = \text{diag}(\gamma_i).
\]

C. Latent Variable based Sparsity Prior Learning

A latent variable based Bayes model is employed to learn the prior parameter and noise level directly from the noisy observation as

\[
p(Y|X,\Theta) = p(Y|X,\Theta) p(\Theta|X) = \prod_i p(Y_i|X_i,\Theta) p(\Theta).
\]

D. CSF Prior based HSIs Denoising

Given the learned prior parameter and the estimated noise level, the sparse representation of the HSI can be obtained by the MAP estimation

\[
Y^{\text{opt}} = \max \{ p(Y|F,\Lambda) \} \quad \text{with } \Lambda = \text{diag}(\Lambda),
\]

Then, the denoised HSI is given as \( X^{\text{opt}} = \Phi Y^{\text{opt}} \).

Experimental Results:

A. Results on synthetic data

- Effectiveness of different potentials

B. Results on real noisy data

- Effect of the cluster number \( K \)

Comparison with other methods

- Comparison with other methods

Contact us: zhanglei211@mail.nwpu.edu.cn
Zoom Better to See Clearer: Human and Object Parsing with Hierarchical Auto-Zoom Net

Fangting Xia¹, Peng Wang², Liang-Chieh Chen³, Alan L. Yuille⁴

University of California, Los Angeles¹,²,³,⁴ Johns Hopkins University⁴

Introduction

- Object Semantic Part Parsing
  - Decompose each object instance into its semantic parts.

- Motivation
  - Estimation of object & part scales from the field-of-view (FOV) of a deep network.

- Semantic part parsing can be more accurate under proper object & part scales.

- Contribution
  - We propose a Hierarchical Auto-Zoom Net (HAZN) to perform scale estimation and part parsing jointly, adapting to the local scales of objects and parts.
  - For small-scale objects, HAZN effectively recovers missing parts and gives clearer boundaries; for large-scale objects, HAZN effectively reduces local part ambiguity.
  - Our approach outperforms the state-of-the-arts by a large margin over the PASCAL part datasets on humans⁴, horses and cows⁴.

Object-Scale Auto-Zoom Net

- Multi-tasking image-level FCN
- Scale estimation network (SEN): regresses the scale and location of an object bounding box (ROIs) for each pixel, together with a confidence map.

- Selection and zooming of object ROIs
- Threshold ROI confidence score and perform non-maximum suppression.
- Estimate the object size for each selected object ROI based on image-level part scores and compute a proper zooming ratio.
- Reuse each selected ROI to a proper standard size.

- Refined part scores by object-level FCN
- Re-estimate the part scores within each zoomed object ROI.
- Merge the results to produce the object-level part score map.

Part-Scale Auto-Zoom Net

- A hierarchical approach
- Add part-scale A2N at the end of object-scale A2N.
- Find objects from the image, and then find parts from each object, followed by part score refinement steps.

- Multi-tasking object-level FCN
- Part parsing network (used in object-scale A2N).
- Scale estimation network (SEN): regresses the scale and location of a part bounding box (ROIs) for each pixel.

- Zooming each part ROI to a standard size
- Refined part scores by part-level FCN
- Re-estimate the part scores within each zoomed part ROI, based on the image and object-level part score.

Merge the results to produce the part-level part score map.

Experiments on Parsing Humans in the Wild

- Dataset
  - We conduct experiments on PASCAL Person-Part⁴, which provides pixel-wise part annotations for every human instance. We merge the part annotations into 6 part classes: head, torso, upper/lower arms, upper/lower legs, plus the background class.

- Comparison with state-of-the-art methods
  - Baselines: DeepLab-LargeFOV⁴, DeepLab-LargeCR⁴ (using CRF as a post-processing step), Multi-Scale Averaging (averaging the results from three fixed scales), and Multi-Scale Attention.
  - Our full model surpasses our baseline network [DeepLab-LargeFOV⁴] by 1.8% in IoU, and we are especially better in terms of flexible parts, such as arms and legs.

- Part parsing accuracy w.r.t. size of human instance

- Extensive visual examples

Experiments on Parsing Animals

- Dataset and evaluation metric
  - We perform experiments on the PASCAL horse cow dataset⁵, which contains part annotations for horse instances and cow instances in PASCAL images.
  - The part parsing performance is evaluated by mean IoU (mIoU), computed as the pixel intersection-over-union (IOU) averages across classes.

- Comparison with state-of-the-art methods
  - Baselines: the semantic part parsing (SP) results¹⁰, HyperColumns (HypC)¹¹, the joint object and part (JOP) results⁶.
  - Our model outperforms the state-of-the-art method (JOP)⁶ by over 5% mIoU, and the improvement is most noticeable for small parts like head and tail.

- Visual examples

Acknowledgements

We would like to gratefully acknowledge support from NSF award CIF-1312726, and NSF STC award CIF-1312616. We also thank NVIDIA for providing us with free GPUs that are used to train deep rendering objectives, many thanks to Jingkai Ao, Zhou Ren, and Kang Chen for proofreading the paper and giving suggestions.

References

Main Idea
The paper tackles RGB-D semantic segmentation by discovering common and modality-specific features from the two modalities.

The loss function of our method is as follows

\[ L = \alpha_{rgb} l_{rgb} + \alpha_{d} l_{d} + \alpha_{c} d(c_{rgb}, c_{d}) - \alpha_{s} d(s_{rgb}, s_{d}). \]  

(1)

where \( l_{rgb} \) and \( l_{d} \) are the pixel-wise losses, \( d(c_{rgb}, c_{d}) \) and \( d(s_{rgb}, s_{d}) \) are the MK-MMD between common and specific features.

MK-MMD
Multiple kernel maximum mean discrepancy (MK-MMD) assesses the similarity between common features and modality specific features. We use the following function to calculate the unbiased estimation of MK-MMD between the common features

\[ d(c_{rgb}, c_{d}) = \frac{n}{n^2} \sum_{i=1}^{n/2} \eta(u_i). \]  

(2)

where \( n \) is the batch size, \( c_{rgb} \) and \( c_{d} \) (1 \( \leq i \leq n \)) are the RGB common feature and depth common feature.

Segmentation accuracy
The 14-class average accuracies of different methods on the NYU V1:

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
<th>Method</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silberman and Fergus [8]</td>
<td>50.0%</td>
<td>Pei et al. [36]</td>
<td>50.5%</td>
</tr>
<tr>
<td>Wang et al. [18]</td>
<td>72.0%</td>
<td>Hermann et al. [17]</td>
<td>50.3%</td>
</tr>
<tr>
<td>KDES-RGB [9]</td>
<td>66.2%</td>
<td>KDES-depth [9]</td>
<td>63.4%</td>
</tr>
<tr>
<td>KDES-RGB-D [9]</td>
<td>71.4%</td>
<td>KDES-Treapath [9]</td>
<td>74.0%</td>
</tr>
<tr>
<td>KDES-MRF [9]</td>
<td>74.0%</td>
<td>KDES-MRF-MRF [9]</td>
<td>78.1%</td>
</tr>
<tr>
<td>B-DeN</td>
<td>76.0%</td>
<td>C-DeN</td>
<td>72.1%</td>
</tr>
<tr>
<td>C-DeN</td>
<td>70.3%</td>
<td>E-DeN</td>
<td>74.1%</td>
</tr>
<tr>
<td>U-DeN</td>
<td>69.0%</td>
<td>Ours</td>
<td>78.8%</td>
</tr>
</tbody>
</table>

The class average accuracies of different methods on the NYU V2:

<table>
<thead>
<tr>
<th>Method</th>
<th>4-class</th>
<th>14-class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Couple [11]</td>
<td>62.5%</td>
<td>36.2%</td>
</tr>
<tr>
<td>Goujou [11]</td>
<td>36.2%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Khan [12]</td>
<td>66.7%</td>
<td>35.1%</td>
</tr>
<tr>
<td>Wang et al. [18]</td>
<td>62.2%</td>
<td>35.1%</td>
</tr>
<tr>
<td>Snuker [15]</td>
<td>65.0%</td>
<td>35.1%</td>
</tr>
<tr>
<td>Hoiem [12]</td>
<td>61.1%</td>
<td>35.1%</td>
</tr>
<tr>
<td>Muller [38]</td>
<td>71.5%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Assumes [37]</td>
<td>72.0%</td>
<td>10.0%</td>
</tr>
<tr>
<td>U-DeN</td>
<td>71.8%</td>
<td>40.2%</td>
</tr>
<tr>
<td>Ours</td>
<td>74.7%</td>
<td>47.3%</td>
</tr>
</tbody>
</table>

Visualization and Result
The following figure visualizes the common features and modality-specific features, which are extracted from deconv 2-2.

For 13-class segmentation, the average class segmentation accuracies of our method are 78.8% on NYU V1 and 52.7% on NYU V2. The following figure shows the segmentation results.
MADMM: A GENERIC NON-SMOOTH OPTIMIZATION ON MANIFOLDS

{ARTIMO.KOVNATSKY, KLAUS.GLASHOFF, MICHAEL.BRONSTEIN}@USI.CH
Faculty of Informatics, Università della Svizzera italiana (USI), Lugano, Switzerland

PROBLEM & CONTRIBUTIONS

Problem

\[ \min_{x \in \mathcal{M}} f(x) + g(Ax), \]

where \( f \) and \( g \) are smooth (e.g., \( \text{tr}(\cdot) \)) and non-smooth (e.g., \( \|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_{\infty} \)) real-valued functions, respectively, \( A \) is a \( k \times m \) matrix, and \( \mathcal{M} \) is a Riemannian manifold-valued function.

Contributions

1. MADMM: the first generic algorithm for non-smooth optimization on manifolds
2. Remarkably simple implementation
3. Applicable to any manifold (read constraint)
4. Straightforwardly modified for a task at hand
5. MADMM converges faster than previous methods in a broad range of applications

MATHEMATICAL BACKGROUND

Manifold Optimization

Method Summary

\[
\begin{align*}
\min_{x \in \mathcal{M}} f(X) + g(AX) & \iff \min_{x \in \mathcal{M}} f(X) + g(Z) \iff \min_{x \in \mathcal{M}} f(X) + g(Z) + \tfrac{\mu}{2} \|AX - Z + U\|_F^2 \\
\text{Initialize } k \leftarrow 1, Z^{(1)} = AX^{(1)}, U^{(1)} = 0. \\
\text{repeat } & \quad X^{(k)} = \text{argmin}_{X \in \mathcal{M}} f(X) + \tfrac{\mu}{2} \|AX - Z^{(k)} + U^{(k)}\|_F^2 \\
& \quad Z^{(k)} = \text{argmin}_{Z \in \mathcal{M}} g(Z) + \tfrac{\mu}{2} \|AX^{(k)} + Z^{(k)} + U^{(k)}\|_F^2 \\
& \quad U^{(k+1)} = Z^{(k)} - AX^{(k)} \\
& \quad k \leftarrow k + 1 \\
\text{until convergence; } & \end{align*}
\]

X-step: a smoothing optimization, only a few iterations are done [1]
Z-step: a proximity operator of \( \tfrac{\mu}{2} g(Z) \) at \( AX + U \), often has a closed form solution (e.g., for \( \|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_{\infty} \))

Acknowledgement

This research was supported by the ERC Starting Grant No.307047 (COMET).

REFERENCES


RESULTS

Compressed modes (\( f = \text{tr}(\cdot), g = \mu \cdot \|\cdot\|_1 \), \( \mathcal{M} = \{X \in \mathbb{R}^{n \times k} : X^T X = I, \text{Stiefel manifold} \})

Ozoliņš et al. [2] proposed a construction of localized Laplacian quasi-eigenvectors by solving

\[
\min_{\Phi \in \mathbb{R}^{n \times k}} \text{tr}(\Phi \Delta \Phi) + \|\Phi\|_F \quad \text{s.t. } \Phi^T \Phi = I,
\]

where \( \Delta \) is a Laplacian represented as \( n \times n \) sparse symmetric matrix, \( \Phi = (\phi_1, \ldots, \phi_k) \) is the \( n \times k \) matrix of the first quasi-eigenvectors arranged as columns, and \( \mu > 0 \) is a parameter, controlling sparseness.

Figure 1. Illustration of the classical MDS approach and the equivalence between Euclidean distance matrices (EDM, D) and positive semi-definite (PSD) \( k \) similarity matrices; \( H = I - \tfrac{1}{n} 11^T \) and \( B^* = X^T X \), where \( X \) is the desired embedding.

Figure 2. Compressed modes problem. From left to right: first compressed modes computed on a human mesh containing \( n = 8k \) points using MADMM with parameter \( c = 10^{-5} \) and three manifold optimization iterations in the X-step; comparison of convergence of different splitting methods and MADMM on a problem on size \( n = 500, k = 10 \) with different random initialization; comparison of convergence of different splitting methods and MADMM on a problem of size \( n = 8k \); convergence of MADMM using different solvers and number of iterations at X-step for the same problem.

Robust Euclidean embedding (REE) (\( f = 0, g = \|\cdot\|_1, \mathcal{M} = \text{fixed-rank positive semi-definite matrices} \))

Cayton and Dasgupta [6] treated an ADMM (notation as in Figure 2):}
Abstract: In existing anisotropic diffusion-based semi-supervised learning approaches, anisotropic graph Laplacian is estimated based on (potentially noisy) function evaluations. We propose to regularize the graph Laplacian estimates. We develop a framework that regularizes the Laplace-Beltrami operators on Riemannian manifolds, and discretize it to a regularizer on diffusion operators on graphs.

Isotropic Laplace-Beltrami operator $\Delta^g$ on a Riemannian manifold $(M, g)$ with a metric $g$ is a second-order differential operator:

$$\Delta^g f = \nabla^g \cdot \nabla^g f,$$

where $\nabla^g$ and $\nabla^g \ast$ are the gradient and divergence operators, respectively. $\Delta^g$ generates the diffusion process on $M$:

$$\frac{\partial f}{\partial t} = -\Delta^g f.$$

Anisotropic Laplace-Beltrami operator $\Delta^D$ is defined based on a symmetric positive definite diffusion operator $D$:

$$\Delta^D f = \nabla^D \cdot \nabla^D f,$$

$D$ controls the strength and direction of diffusion at each point $x$ on $M$.

Regularizing $\Delta^g$ by regularizing $D$ as a surrogate:

1) Kernel-based $\Delta^g$ representation [HALO5]: A consistent kernel-based estimate $\Delta^g_k f$:

$$\Delta^g_k f(x) = \frac{1}{|E(k)|} \sum_{(i,k) \in E(k)} [A_k(x,f)](i,j),$$

where $A_k(x,f)$ is the kernel matrix, and $k(x, y) = \exp(-d(x,y)/\sigma)$ and $\sigma$ being the embedding of $M$ into $\mathbb{R}^n$.

2) Equivalence of metric and diffusivity operator on manifolds:

Proposition 1 (KTP15): The anisotropic Laplace-Beltrami operator $\Delta^D$ on a compact Riemannian manifold $(M, g)$ is equivalent to the Laplace-Beltrami operator $\Delta^g$ on $(M, \nabla^g)$ with a new metric $\nabla^g$ depending on $D$.

When the diffusion operator $D$ is uniformly positive definite, $\nabla^g$ is explicitly obtained as $\nabla^g(x) = g(x)D^{-1}(x)$, where $g(x)$ and $D(x)$ are the coordinate matrices of $g$ and $D$ at each point $x$, and $\nabla^g(x) = \sqrt{\det(g^T)} \sqrt{\det(D)}^{-1}.$

Anisotropic diffusion on $(M, g)$ is equivalent to diffusion on $(M, \nabla^g)$ with a new metric $\nabla^g$. We regularize the Laplacian on manifolds by enforcing the smoothness of the corresponding diffusion operator. Instantiating this for graphs, our algorithm enforces the smoothness of the graph diffusion functions $D_1 \cdots D_t$.

Semi-supervised learning results: The three best results for each dataset are ranked with boldface blue, plain green, and plain orange fonts, respectively. LNP [WAzo6] requires explicitly calculating the Euclidean distances between data points, and so it cannot be directly applied to MPEG7 and Swd (Swedish leaf) datasets. The final Avg. % column shows the mean percentage difference from the best result across all datasets, where 100% would indicate that particular technique was best across all datasets.

Acknowledgement: Kwang In Kim thanks EPSRC EP/M00533X/1.

Reference:


Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classification

Maxime Bucher1,2, Stéphane Herbin1, Frédéric Jurie2
ONERA - The French Aerospace Lab1, GREYC2

Motivation

- **Image classification** goal is to predict the class of an object present in an image.
- Traditional classification pipeline design → train a classifier from annotated image database.
- However, in many applications having access to image data is often difficult.

- **Zero-Shot learning**: recognize object never seen during training using attribute descriptions as an intermediate knowledge-based semantic image representation.

  → **Problem**: attribute embedding space can be redundant and noisy.

- **Our approach**: use a Metric Learning framework with dimension reduction and space transformation → improve semantic embedding.

Main idea

- **Attribute embedding**: intermediate representation level, understandable by human designers and sufficiently formal to be the support of algorithmic inferences.
- **Attribute representation weaknesses**: may not be the ideal embedding space, can be too redundant and noisy to support reliable inferences.

Our goal → improve semantic embedding by learning a **consistency score** \( S(X, Y) \) between image \( X \) and attribute vector \( Y \).

Our idea → optimize jointly the attribute embedding and the classification metric, in a multi-objective framework:

- **Attribute detection**: improve image embedding capacity
- **Metric Learning**: optimization based on asymmetric positive/negative pairs of attributes/images, acts as a discriminating space transformation and dimension reduction.

Consistency score:

\[
S(X, Y) = \|X^T W_A - Y^T W_A\|_2
\]

Optimization criterion:

\[
\mathcal{L}(W_A, W_Y) = \max_i \left(0, 1 - Z_i (\tau - \|X^T W_A - Y^T W_A\|_2^2) + \lambda \sum_i \|Y_i - X^T W_A\|_2^2 + \text{Regularization}\right)
\]

Experiments

At test time the learned model can predict the consistency of a test image with a given set of attributes:

- every unseen class is described by an attribute vector
- the consistency score between the tested image and all class vectors is computed
- the label with maximal score is the predicted class

State of the art performances in terms of accuracy on 4 Zero-shot datasets:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Method</th>
<th>aP&amp;Y</th>
<th>AwA</th>
<th>CUB</th>
<th>SUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lampert et al. [1]</td>
<td>38.16</td>
<td>57.23</td>
<td>-</td>
<td>72.00</td>
<td></td>
</tr>
<tr>
<td>Romera-Paredes et al. [2]</td>
<td>24.22±2.89</td>
<td>75.32±2.28</td>
<td>-</td>
<td>82.10±0.32</td>
<td></td>
</tr>
<tr>
<td>Zhang et al. [5]</td>
<td>46.23±0.53</td>
<td>76.33±0.83</td>
<td>30.41±0.20</td>
<td>82.50±1.32</td>
<td></td>
</tr>
<tr>
<td>Zhang et al. [4]</td>
<td>50.35±2.97</td>
<td>80.46±0.53</td>
<td>42.11±0.55</td>
<td>83.83±0.29</td>
<td></td>
</tr>
<tr>
<td>Ours w/ML</td>
<td>47.25±0.48</td>
<td>73.81±0.13</td>
<td>33.51±0.98</td>
<td>74.91±0.12</td>
<td></td>
</tr>
<tr>
<td>Ours w/o constraint</td>
<td>48.47±1.24</td>
<td>75.69±0.56</td>
<td>38.35±0.49</td>
<td>79.21±0.87</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>53.15±0.88</td>
<td>77.32±1.03</td>
<td>43.29±0.38</td>
<td>84.41±0.71</td>
<td></td>
</tr>
</tbody>
</table>

→ **Few-shot Learning** and **image retrieval** also possible based on the metric learning based consistency score (see article).

Classical VS Zero/Few-shot Learning

- **Classical classification**: class inference decision based on training images
- **Zero/Few-shot learning**: use an attribute representation to compensate for the lack of training image data

Metric Learning

Basic idea: learn a distance function that assigns small (resp. large) distance to pairs of examples that are semantically similar (resp. dissimilar).

Image - attribute samples

References

A Sequential Approach to 3D Human Pose Estimation: Separation of Localization and Identification of Body Joints

Ho Yub Jung*, Yumin Suh, Gyeongsik Moon, and Kyoung Mu Lee
Department of ECE, ASRI, Seoul National University, Seoul, Korea
*Div. of CESE, Hankuk University of Foreign Studies, Korea

INTRODUCTION & BACKGROUND

Human Pose from Depth Image
- Finding body joint positions from a single depth image

Challenges & Motivation
- Localization problem: The 15 3D position of each joint must be accurately estimated for human pose.
- Identification Problem: Joint positions must be simultaneously identified during localization. This is especially hard for left and right limb distinction

Our solution
- Solve the localization problem first
- All joint positions (w/o joint labels) have sufficient information to provide body part identification

Contributions
- New machine-learning based approach that separate human pose problem into localization and identification problem
- A state-of-the-art human pose estimation result

STAGE1: Localization Problem

- Regression tree used
  - Single regression tree is trained to output the offset to the nearest joint position.
- How to inference?
  - All joint positions (without labels) estimated using simple K-means clustering.

STAGE2: Joint Identification Problem

- Localization is sufficient
  - 15 joint positions (without joint labels) have sufficient information to provide joint identification as shown in the figure.
- How to identify each joint?
  - A simple nearest distance exemplar retrieval system is shown to provide a good joint identification.

EXPERIMENTS & CONCLUSION

Comparison with state-of-the-art
- Ganapathi et al., ECCV 2012
- Ye and Yang, CVPR 2014
- Jung et al. CVPR 2015

EVAL DB (within 10 cm)

<table>
<thead>
<tr>
<th>Methods</th>
<th>Ours</th>
<th>Ye&amp;Yang 14</th>
<th>Jung 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>mAP (%)</td>
<td>98.3</td>
<td>92.1</td>
<td>92.6</td>
</tr>
</tbody>
</table>

Poses Estimation Examples

Regression Tree Localization and Joint Identification

Conclusion
1. Separation of localization and identification gives a large benefit.
2. Localization is enough to estimate joint locations.
3. Our method outperforms other methods by a large margin.
A Novel Tiny Object Recognition Algorithm Based on Unit Statistical Curvature Feature

Yimei Kang, Xiang Li
College of Software, Beihang University, Beijing, China

Problem
Recognize tiny objects whose sizes are in the range of 15 × 15 to 40 × 40 pixels.

Contribution
- A novel image feature descriptor, unit statistical curvature feature (USCF), is proposed based on the statistics of unit curvature distribution to represent the local general invariant features of the image texture.
- USCF algorithm had high recognition rate for object images in any size including tiny object images.
- USCF is invariant to rotation and linear illumination variation, and is partially invariant to viewpoint variation.

Experiments
The recognition rate of USCF algorithm was the highest for tiny object recognition compared to SIFT [1]. SURF [2], ORB [3], gray histogram, entropy, unit entropy, GIST [4], HoG [5] and Hu's moment invariants algorithms on ALOI-COL Database, COLI-100 Database, ETH-80 Database and ETHZ another 53 Objects Database, respectively under complex test conditions with simultaneous rotation, illumination, viewpoint variation and background interference.

Application
Recognize tiny objects in a distance in real time.

Proposed Algorithm USCF
1. Employ least square method to fit curved surfaces of object images:
   \[
   f^l(x, y) = \sum_{i=0}^{9} \sum_{j=0}^{9} c_{ij} x^i y^j \]
   \[
   f^l(x, y) = f^l(x, y) - z_{ij} \]
   \[
   \frac{partial}{partial f^l} = 0
   \]
2. Calculate the Gaussian curvature and mean curvature of each pixel images according to the fitted curved surfaces:
   \[
   H = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1+|f_x|^2)^2}, \quad K = \frac{f_{xx} f_{yy} - f_{xy}^2}{(1+|f_x|^2)^2} \]
3. Build the curvature feature space in \( \mathbb{R}^{O_u} \) and non-uniformly partition it into \( w \times v \) units according to the curvature distribution density.
   \[
   Area_{ij} = \left[ (H, K) | H_{ij-1} < H < H_{ij+1}, K_{ij-1} < K < K_{ij} \right]
   \]
4. Count the number of pixels in each unit to generate the USCF matrix of an image.
   \[
   D = \left[ count(Area_{ij}) \right]_{w \times v}
   \]
5. Match the similarity of the USCF matrices of the candidate image and the template image with Euclidean distance.
   \[
   dist = \sum_{i=1}^{w-1} \sum_{j=1}^{v-1} (D_i,j - \sum_{k=1}^{w} D_k, j) (\sum_{k=1}^{v} D_i, k)^{-1}
   \]

Parameter Selection in Experiments
- The experimental images were shrunk from the images in the databases to the sizes of 15 × 15, 20 × 20, 25 × 25, 30 × 30, 35 × 35 and 40 × 40 pixels, by using bicubic interpolation, respectively.
- The distribution area of \( K \) and \( H \) are partitioned into 17 parts and 11 parts as following, respectively.
  \[
  H_i = \{i - 5.5 | j / 11 \} \times 10^{-1.5 \pm 2}, \quad j = 0, 1, ..., 11
  \]
  \[
  K_i = \{i - 8.5 | j / 7 \} \times 10^{-1.5 \pm 2}, \quad j = 0, 1, ..., 7
  \]
- For unit entropy and HoG algorithms, \( 5 \times 5 \) pixels is used as the size of each unit.

Robustness against Rotation
- COLI-100 Database was used to evaluate the anti-rotation performance.
- The candidate object images were rotated clockwise by 15, 60, 90 and 175 degrees, respectively.

Comparison of the Test algorithms on Images from Videos

Comparison of the Test algorithms on Images from Videos

Table 1: The information of the test images used for the experiment

<table>
<thead>
<tr>
<th>Video</th>
<th>Model</th>
<th>Shape</th>
<th>Scale</th>
<th>Background</th>
<th>Illumination</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALOI-COL</td>
<td>800</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>ETH-80</td>
<td>800</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>ETHZ</td>
<td>800</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2: The performance of the test algorithms on the test images

<table>
<thead>
<tr>
<th>Video</th>
<th>Model</th>
<th>Shape</th>
<th>Scale</th>
<th>Background</th>
<th>Illumination</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALOI-COL</td>
<td>800</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>ETH-80</td>
<td>800</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>ETHZ</td>
<td>800</td>
<td>80</td>
<td>80</td>
<td>20</td>
<td>100</td>
</tr>
</tbody>
</table>

Conclusions
A novel object recognition algorithm, USCF algorithm, based on unit statistical curvature feature was proposed. The experimental results showed that USCF algorithm is robust to rotation and illumination variation, and can tolerate slight viewpoint variation. Under complex test conditions with simultaneous rotation, illumination, viewpoint variation and background interference, the recognition rate of USCF was the highest among all ten tested algorithms. USCF cost less than 40 ms on a desktop PC with Intel(R) Core(TM) i5-3470 CPU when the image sizes were smaller than 40 × 40 pixels, which indicates that USCF can be applied in a real time application for tiny object recognition.

References
**Head Reconstruction from Internet Photos**

Shu Liang, Linda G. Shapiro, Ira Kemelmacher-Shlizerman
University of Washington, Computer Science & Engineering

**ECCV 2016**

---

**Results**

<table>
<thead>
<tr>
<th>Pose</th>
<th>-90</th>
<th>-60</th>
<th>-30</th>
<th>0</th>
<th>30</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bush</td>
<td>185</td>
<td>62</td>
<td>118</td>
<td>371</td>
<td>113</td>
<td>80</td>
<td>191</td>
</tr>
<tr>
<td>Putin</td>
<td>131</td>
<td>58</td>
<td>151</td>
<td>413</td>
<td>121</td>
<td>61</td>
<td>151</td>
</tr>
<tr>
<td>Obama</td>
<td>65</td>
<td>51</td>
<td>126</td>
<td>284</td>
<td>177</td>
<td>55</td>
<td>75</td>
</tr>
<tr>
<td>Clinton</td>
<td>115</td>
<td>47</td>
<td>114</td>
<td>332</td>
<td>109</td>
<td>61</td>
<td>66</td>
</tr>
</tbody>
</table>

Number of Photos we used in each pose cluster

---

**Comparison**

![Image of comparison](image)

**Limitations**

We have shown the first results of head reconstructions from Internet photos, but:
1) Lambertian model doesn’t capture hair well. We also haven’t worked on reconstructing details.
2) This model could be combined with template based models.
3) 180 degrees for side views were labeled manually
4) We have not reconstructed a complete model; the top of the head is missing. To solve this we would need to add photos with different elevation angles.

---

**2. Photometric Stereo**

Frontal view photos to n x p matrix Q.
$n$: number of photos, $p$: number of pixels of the facial mask. Rank-4 PCA is computed to factorize into lighting and normal: $Q = LN$, with ambiguity, Q=LA’AN.

Resolve the Generalized Bas-Relief (GBR) ambiguity using a template 3D face of a different individual, i.e.,

$$\min ||N_{\text{template}} - AN_{\text{face}}||^2$$

The surface normals are integrated to create $D_0$ by solving linear equations that satisfy gradient constraints:

$$n_0(D_{x+1} - D_{x}) = n_x$$
$$n_0(D_{y+1} - D_{y}) = n_y$$
$$n_0(D_{z+1} - D_{z}) = n_z$$

This generates a sparse 2p x 2p matrix $M$, and we can solve for:

$$\text{argmin}_{M} ||M - I||^2$$

---

**3. Boundary-Value Growing**

![Image of boundary-value growing](image)

**Limitations**

Observation: each view cluster has one particularly well-reconstructed area

---

**References**


---

**Dataset**

The Conditional Lucas & Kanade Algorithm

Chen-Hsuan Lin, Rui Zhu, and Simon Lucey
{chenhsul, rz1}@andrew.cmu.edu, slucey@cs.cmu.edu

Contributions
- Establish theoretical connections: The Lucas & Kanade Algorithm (LK) ↔ Supervised Descent Method (SDM)
- The Conditional LK Algorithm: efficient aligner which achieves comparable performance with little training data

LK
- Aligns a source image \( I \) against a template image \( T \) by minimizing their appearance error
\[
\min_{\Delta p} \left\| I(p) - T(\Delta p) \right\|^2_2 \quad \text{(inverse-compositional)}
\]
- 1st-order Taylor approximation
\[
\min_{\Delta p} \left\| I(p) - T(0) - \nabla T(0) \frac{\partial \phi(x; 0)}{\partial p} \Delta p \right\|^2_2
\]
- Solve
\[
\Delta p = \left( \frac{\partial \phi(x; 0)}{\partial p} \right)^{-1} \left( I(p) - T(0) \right)
\]

SDM
- Learns the appearance-geometry linear relationship from synthetically generated data \( S = \{\Delta p_n, I_n(p_n \circ \Delta p_n)\}_{n=1}^N \)
- Linear regressors are trained from independently sampled data per iteration

Training:
\[
\min_{R} \sum_{n=1}^N \left\| \Delta p_n - R[I_n(p_n \circ \Delta p_n) - T(0)] \right\|^2_2 + \Omega(R)
\]
- Prediction of the geometric displacement is learned conditioned on the appearance

Evaluation:
\[
\Delta p = R[I(p) - T(0)]
\]

Similarity
- Both assume a linear relationship between appearance and geometry
- Solve for geometric updates iteratively until convergence is reached

Difference
LK
- Pixel independence assumption
- Generative appearance synthesis

SDM
- Full dependency across pixels
- Conditional learning objective

Conditional LK
- Learn the image gradients \( \nabla T(0) \) conditioned on the appearance
- Conditional learning objective
- Pixel independence assumption
- Warp swapping property:
  Geometric warp functions can be swapped and combined with the conditional image gradients \( \nabla T(0) \) to form another series of linear regressors
- Non-linear least squares problem

Experiments
Visualization of \( \nabla T(0) \)

Convergence Analysis
Frequency of convergence

Warp Swapping

Applications
Low frame-rate tracking

Simultaneous Results for different warpings, feature images, examples/iteration, and test perturbations
BLUE: IC-LK
YELLOW: Conditional LK
Where should saliency models look next?

Zoya Bylinskii, Adrià Recasens, Ali Borji, Aude Oliva, Antonio Torralba, Frédéric Durand

Have saliency models begun to converge on human performance? We re-examine the current state-of-the-art using a fine-grained analysis on image types, individual images, and image regions. We quantify up to 60% of remaining errors of saliency models. To continue to approach human-level performance, saliency models will need to discover higher-level concepts in images and reason about the relative importance of image regions.

All state-of-the-art models are neural networks. Spikes in performances are observable on all metrics. Metrics like NSS and IG are more informative than others.

Images most representative of model performance

Imbuing predicted saliency maps with ground truth

Replacing saliency predictions in regions of interest with ground truth can approximate performance gains on MIT Saliency Benchmark.

Finer-grained datasets and metrics

Finer-grained datasets can break up model performance by image category and uncover performance gaps.

Assigning correct relative importance to faces

Under-predicted Missed depictions Over-predicted

Which face is most important?

Current saliency models are good face detectors. The next challenge is analyzing the relative importance of faces compared to other faces and image content.

Which is the most important piece of text?

Which text in a scene provides the most relevant information for image understanding? At which point does saliency modeling become user-specific instead of populations-specific?

An explicit model of gaze can provide important cues not currently used by saliency models (above). In a similar manner, body posture and hand positions can point to objects of interest in a scene (left).
Robust Face Alignment Using a Mixture of Invariant Experts

Oncel Tuzel¹, Tim K. Marks¹, Salil Tambe²
¹Mitsubishi Electric Research Laboratories (MERL), ²Intel

1. Introduction

Goal
- Face alignment (finding the locations of a set of facial landmark points) with accuracy that improves upon the state of the art.

Problem
- Difficult due to variations in pose, facial expressions, illumination, and occlusion.
- SDM [17] uses a cascade of regression functions that operate on extracted SIFT features to iteratively estimate facial landmark locations. But having just one model for all faces forces the SDM regression functions to be too generic, limiting accuracy when there are large variations in expressions and pose.

Solution
- We propose a Mixture of Invariant Experts (MIX) to represent these variations. Each expert learns a regression model tuned to align faces in a particular range of poses and expressions.

Contributions
1. A mixture-of-experts regression at each stage of the cascade. Each expert regression function is specialized to align a different subset of the input data (a particular range of expressions and poses).
2. A transformation invariant step, before each stage of regression, that makes our method invariant to a specified class of transformations (e.g., 2D affine).
3. A novel transformation-invariant clustering algorithm to learn the prototype shapes used in the mixture model.
4. A simple extension to the feature vectors that penalizes deviations of feature locations from a prototype face shape.

Each expert corresponds to a prototype shape, whose landmark locations are the center of a learned affine-invariant cluster.

Weights are determined based on the current estimate of the landmark locations.

2. Approach

Transformation-Invariant SDM (TI-SDM): We transform the image along with the current estimate of landmark locations to a prototype frame, then apply the regression model to get an updated estimate. We then transform the updated estimate back to the image frame.

Algorithm 1 Stage k of Transformation-Invariant SDM

Inputs: Prototype shape, Regression function (\(W_k, b_k\)).

1. Apply affine transformation \(A_k\) that maps \(x_o\) to \(\tilde{x}_k\).
2. Warp prototype features \(\tilde{f}_k\) to \(f_k(\tilde{x}_k)\).
3. Extract features: \(f'(\tilde{x}_k) = f_k(\tilde{x}_k)\).
4. Linear regression: \(\tilde{x}_{k+1} = \tilde{x}_k + b_k f'_{k+1}\).
5. Warp back to image: \(x_{k+1} = A_k^{-1}(\tilde{x}_{k+1})\).

Output: Landmark Locations \(x_{k+1}\).

MIX: At each stage \(k\) \(\in\), we apply TI-SDM for each of the \(i\) experts. The regression output is a weighted average of the experts’ regression outputs.

Algorithm 2 Mixture of Invariant Experts (MIX)

Inputs: Initial landmark estimate \(x_i\), Experts’ \(\tilde{f}_k\), Initial landmark estimate \(x_i\), Experts’ \(\tilde{f}_k\).

1. for \(k = 1\) to \(K\) do
2. Compute warp function \(w_k\) for each expert.
3. Compute lift warping \(w_k(x)\).
4. Apply lift warping to \(x_i\).
5. end for
6. Average over \(L\) experts: \(x_{i+1} = \frac{1}{L} \sum_{l=1}^{L} x_i\).
7. end for

Output: Final landmark locations \(x_{i+1}\).

3. Training

- We use the training samples from the LFPW and HELEN datasets, with annotations from the 300W dataset. Our regression cascade has 3–4 stages of coarse alignment, followed by 1–2 stages of fine alignment.
- Initial coarse initialization: Random PCA perturbation, rotation, translation, scaling, and i.i.d. Gaussian noise.
- Fine initialization: Small amount of i.i.d. Gaussian noise.
- Each regression function \(W_k, b_k\) is successively learned by solving the Tikhonov regularized L2 loss function:

\[
(W_k^i, b_k^i) = \arg \min_{W,b} \sum_{i=1}^{M} \alpha_i \left( \|\Delta(x_k^i) - Wf(I, x_k^i) - b\|_2^2 + \gamma \left(\|W\|_F^2 + \|b\|_2^2\right) \right)
\]

The soft assignments are computed using \(\alpha_i(x) = e^{-\frac{1}{2}(x-x_i)^T(x-x_i)} / \sum_{i=1}^{M} e^{-\frac{1}{2}(x-x_i)^T(x-x_i)}\), where \(x^i = \arg \min_{A_k} \|A_k(x) - x_i\|_2^2\).

4. Results

Our MIX algorithm outperforms SDM in cases with large out-of-plane rotations, occlusions, and unusual facial features (e.g., facial hair).

Normalized Area Under Curves (shown below) on 300W (Combined) dataset.

5. References

1. Short Summary

**Contribution**

**Advantages**
- Informative descent direction
  - computed using the marginals over each output class, similar to Exponentiated-Gradient.
- Optimal step-size in the descent direction
  - that guarantees an increase in the dual objective, similar to Frank-Wolfe.
- Block coordinate formulation
  - similar to the one proposed for Frank-Wolfe, which allows us to solve large-scale problems.

2. Multi-class SVM

- **Input:** x ∈ X; **Output:** y ∈ Y, where Y = {1,..., c} & c is the no. of classes
- **Feature representation for sample:** φ(x)
- **Joint Feature map for sample and candidate class:**
  \[ \phi(x, y) = \sum_{j=1}^{c} \psi_j \mathbf{v}_j^T \phi(x, y) \] where, \( \psi_j = \{ \phi(x) \text{ if } j = y; \ 0 \text{ otherwise}. \)

- **Prediction:**
  \[ y_{opt} = \arg \max_y w^T \phi(x, y) \]

- **Learning:**
  \[ \min_{\alpha} 2 \sum_{n=1}^{N} \frac{1}{2} \|x_n - A \alpha_n \|^2 + \sum_{n=1}^{N} \sum_{i=1}^{c} \xi_i \]
  s.t. \( \forall i, j : w^T \phi(x, y) - w^T \phi(x, y) \geq \xi_{ij} \)

3. Partial Linearization framework

- **Consider Optimization problem:** \( \min_{\alpha} T(\alpha) \)
  - Where, \( T(\alpha) \) is a convex and continuously differentiable function and \( U \) is a compact & convex set.

4. Partial Linearization based Optimization for Multi-class SVM

- **Appropriate design of Surrogate function:**
  \[ f(\alpha, \alpha') = \frac{1}{2} \sum_{n=1}^{N} \sum_{i=1}^{c} \alpha_i \log(\alpha_i) \]
  - Resulting problem is efficiently optimizable
  - Captures information about violation of each primal constraint

- **Corresponding update direction:**
  \[ s_i = \exp\left(\log(\alpha_i') - \frac{1}{2} (\lambda (y_i - y_{opt_i})) \right) \]

- **Block Coordinate Partial Linearization**
  - **For Temperature:** \( T = 0, \) **Partial-Linearization**
  - **For Temperature:** \( T = 0, \) **Frank-Wolfe**
  - **For Step-size:** \( \gamma = 1, \)** **Partial-Linearization**
  - **For Step-size:** \( \gamma = 1, \)** **Exponentiated-Gradient**

- **Partial Linearization based optimization for Structured-SVM**
  - Can be used for tree-structured output spaces.
  - Message passing for computing exact marginals for the update step.

5. Experiments and Results

- **Problems and Datasets:** Training Multi-Class SVM or Structured SVM models for Vision tasks like Classification (Pascal VOC), Object Recognition (CIFAR-10, Pascal VOC), Gesture Recognition (MSRC-12) and Hand written word recognition (OCR).

- **Methods:** We compared our Block-Coordinate Partial Linearization algorithm (BCPL) with Block-coordinate Frank-Wolfe (BCFW) algorithm [2] and Online exponentiated gradient (OEG) algorithm.

- **Results:**
  - Example training curves (Dual objective Vs. time (sec)):

- **Mean training time (sec) of Multi-Class SVM or Structured-SVM models for different tasks:**

6. References

Introduction

• Depth estimation from a single monocular image is a fundamental problem in computer vision, with various applications in stereo vision, robotics, and scene understanding.
• Unlike the traditional approaches that learn a regressor from image to depth indirectly, we first perform joint dictionary learning to bridge the similarity gap between 2D image patches and 3D depth maps to facilitate cross-modal retrieval.
• Comparing to deep learning method, the proposed method maps the query image to the dataset images directly, which can hardly over-fit the dataset and allow us to reuse our model in any situation without re-training.

Model Learning

Cross-Modality based Prior Depth Inference:
\[
\min \|X_{\text{im}} - D_{\text{im}} Y\|^2_p + \alpha \|X_{\text{dep}} - D_{\text{dep}} Y\|^2_p \tag{2}
\]

Large Margin Structure Inference:
\[
\psi_D(I_{\text{dep}}, I_{\text{im}}, I_{\text{im}}) = \phi_D(I_{\text{im}}) \psi_{\text{dep}}(I_{\text{dep}}, I_{\text{dep}}, I_{\text{im}}, I_{\text{im}}) \psi_{\text{dep}}(I_{\text{dep}}, I_{\text{im}}) \tag{16}
\]

The Algorithm 1: the Proposed Method

Input
Query Image \( I_{\text{im}} \) \( \in \mathbb{R}^{w \times h} \)
Corresponding Candidate 3D Models \( I_{\text{dep}} \) \( \in \mathbb{R}^{w \times h} \) \( (i = 1, \ldots, N) \)

1. Cross-Modality based Prior Depth Inference
   (a) Extract overlapped image patches \( X_{\text{im}} \) \( \in \mathbb{R}^{w \times h} \) \( (i = 1, \ldots, n) \) from \( I_{\text{im}} \) and corresponding depth map patches \( X_{\text{dep}} \) \( \in \mathbb{R}^{w \times h} \) \( (j = 1, \ldots, m) \) from \( I_{\text{dep}} \) using Eq.2 to calculate \( D_{\text{im}} \) \( \in \mathbb{R}^{w \times h} \) and \( D_{\text{dep}} \) \( \in \mathbb{R}^{w \times h} \) \( (i = 1, \ldots, N) \) using Eq.2.
   (b) Extract non-overlapped image patches \( X_{\text{im}} \) \( \in \mathbb{R}^{w \times h} \) \( (k = 1, \ldots, m) \) from \( I_{\text{im}} \) using Eq.2 to calculate the corresponding initial depth map patches \( X_{\text{dep}} \) \( \in \mathbb{R}^{w \times h} \) \( (i = 1, \ldots, n) \).
   (c) Obtain the prior depth \( I_{\text{dep}}^0 \) of the entire image \( I_{\text{im}} \).

2. Large Margin Structure Inference
To minimize Eq. 16, is equivalent to minimize
\[
\begin{align*}
\min & \psi_D(I_{\text{dep}}, I_{\text{im}}, I_{\text{im}}) = & \min \psi_{\text{dep}}(I_{\text{dep}}) + \psi_{\text{im}}(I_{\text{im}}, I_{\text{im}}) \\
& + \psi_{\text{dep}}(I_{\text{dep}}, I_{\text{im}}, I_{\text{im}}) \tag{9}
\end{align*}
\]

Eq. 16 can be transformed into the following format
\[
\begin{align*}
\min & \psi_D(I_{\text{dep}}, I_{\text{im}}, I_{\text{im}}) = \sum A_{i,j} \mathbf{b}_i - \mathbf{b}_j \tag{10}
\end{align*}
\]

To minimize Eq. 10, we can get the \( \mathbf{b}_j \) iteration solution of \( I_{\text{im}} \) by gradient descent
\[
I_{\text{im}} = \left( \sum A_{i,j} \mathbf{b}_i - \mathbf{b}_j \right) + \varepsilon \left( \sum A_{i,j} \mathbf{b}_i - \mathbf{b}_j \right)^2 + \varepsilon \tag{11}
\]

where \( A_{i,j} \) is the \( i,j \)th row of \( A_t \), \( b_i \) is the \( i \)th element of vector \( b_t \) and \( \varepsilon = 10^{-10} \)

Output
The optimized depth map \( I_{\text{im}}^0 \) of image \( I_{\text{im}} \).

Experiments

Table 1. Result comparisons on the Make3D dataset (C2) Errors are computed in the regions with ground-truth depth less than 70.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error(C1) (lower is better)</th>
<th>Error(C2) (lower is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rel Ig10 mm</td>
<td>rel Ig10 mm</td>
<td></td>
</tr>
<tr>
<td>Make3D [26]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Depth Fusion [12]</td>
<td>0.386</td>
<td>0.158</td>
</tr>
<tr>
<td>Depth Fusion作风 [13]</td>
<td>0.378</td>
<td>0.137</td>
</tr>
<tr>
<td>Depth Transfer [11]</td>
<td>0.350</td>
<td>0.131</td>
</tr>
<tr>
<td>Ours</td>
<td>0.342</td>
<td>0.130</td>
</tr>
</tbody>
</table>

*Results reported in Depth Transfer [11].

Fig. 1. Visualization results of Dictionaries: The left four columns are visualized from Make3D dataset and the right four columns from NYUv2 dataset. The first row consists of test images, the second row and third row consists of RGB format feature dictionaries and depth dictionaries, respectively, which are trained by the candidate (22) images.

Table 2. Result comparisons on the Make3D dataset without MRF to fine-tune.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error(C1) (lower is better)</th>
<th>Error(C2) (lower is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rel Ig10 mm</td>
<td>rel Ig10 mm</td>
<td></td>
</tr>
<tr>
<td>Make3D [26]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Depth Transfer [11]</td>
<td>0.396</td>
<td>0.176</td>
</tr>
<tr>
<td>Ours</td>
<td>0.386</td>
<td>0.158</td>
</tr>
</tbody>
</table>

*Results reported in Depth Transfer [11].

Fig. 3. Examples of depth predictions on the Make3D dataset.

Table 5. Result comparisons on the NYUv2 dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Error (lower is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rel Ig10 mm</td>
<td>rel Ig10 mm</td>
</tr>
<tr>
<td>Make3D [26]</td>
<td>-</td>
</tr>
<tr>
<td>Depth Fusion [12]</td>
<td>-</td>
</tr>
<tr>
<td>Depth Fusion作风 [13]</td>
<td>-</td>
</tr>
<tr>
<td>Depth Transfer [11]</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>-</td>
</tr>
</tbody>
</table>

*Results reported in Depth Transfer [11].

Acknowledgement

This work was supported in part by the Major State Basic Research Development Program of China (973 Program) under Grant 2015CB351804, the National Key R&D Program (No. 2016YFB1001503), the National Science Foundation of China (No. 61222100, No. 61733076, No. 61472101, No. 61402388 and No. 61572410), the CCF-Tencent Open Research Fund, and the Open Projects Program of National Laboratory of Pattern Recognition.
Approximate OrthoNormal (AON) Regularizer

\[ ||WW^T - I||^2. \]  

- AON is simple and computationally efficient.
- AON achieves robustness and better generalization by preventing rank-deficient mappings.

\[ \text{ApproximateOrthoNormal (AON) Regularizer} \]

**Overview**

- WARCA is a new model for large-scale learning of Mahalanobis distances.
- WARCA optimizes the precision at top ranks by combining the WARP loss with a regularizer that favors orthonormal linear mappings and avoids rank-deficient embeddings.
- Non-linear WARCA through kernel-trick when data-set size permits kernel computation.
- Benchmarks on nine re-identification data-sets shows state-of-the-art performance both in terms of accuracy and speed.
- Experimental analysis also show how our new regularizer improves the performance.

\[ |S| \]

\[ \sum_{(i,j) \in S} L(\text{rank}_{ij}(F_W)) \]

**Approximate OrthoNormal (AON) Regularizer**

\[ \text{Max-margin Reformulation} \]

\[ \arg\min_{W} \frac{\lambda}{2} ||WW^T - I||^2 + \frac{1}{|S|} \sum_{(i,j) \in S} L(\text{rank}_{ij}(F_W)) \]  

where:

\[ \xi_{ij} = F_W(x_i, x_j) - F_W(x_i, x_k) \]  

and \( \text{rank}_{ij}(F_W) = \sum_{k \in T_i} I_{F_W(x_i, x_k) \leq F_W(x_i, x_j)} \)  

- Efficiently optimized using SGD.

**Experimental Results**

**CMC Curves**

**Analysis of the AON regularizer**

**Analysis of the Training Time**

Code available at: [https://github.com/idiap/warca](https://github.com/idiap/warca)
Overview

Problem. Dense 3D reconstruction of a dynamic foreground subject from a pair of unsynchronized videos with unknown temporal overlap.

Challenges:
1. How to identify temporal overlap in terms of estimated dynamic geometry.
2. How to robustly estimate geometry without knowledge of temporal overlap.

Key Ideas:
1. Define the cardinality of the maximal set of locally rigid feature tracks as a measure of spatio-temporal consistency of a pair of video sub-sequences.
2. Develop a closed-loop track correspondence refinement process to find the maximal set of rigid tracks.

Contributions:
1. We exploit the correlation between temporal alignment errors and geometric estimation errors.
2. We provide a joint solution to the geometry estimation and temporal video alignment problems.
3. Model-free (i.e. data-driven) framework with wide applicability.

Local Rigidity Test

\[ \sum_{i=2}^{W} |x_{m,i-1} - x_{n,i-1}|^2 - |x_{m,i} - x_{n,i}|^2, \quad T_{m}, T_{n} \in C_{d}, \quad T_{1} = \{X_{HE}\} \]

0. The correspondences are computed along the epipolar lines between source and target video frames.
1. If the initial correspondences are within the same local rigid region, distance between 3D points didn’t change over a time interval.
2. For correspondences located in different rigid regions, distances between 3D points change over a time interval.
3. For 2D feature points haven’t passed rigidity test, we iteratively change it’s correspondence (pick other positions on the epipolar line) and do the local rigidity until it pass the test or all candidate positions are tried.

Outlier Detection within Local Rigid Regions

(a) outliers on the left leg are detected because they located in different rigid parts.
(b) outliers on the right waist (same rigid part as left waist) are removed because they are far away from majority of the other trajectories
(c) correct correspondences are the minority (there might be repetitive correspondences in the target frame).

Experiments

Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th># Video Frames/GT 3D Points/Synchronized</th>
<th>Moving Cameras/Outdoor Scene</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETH</td>
<td>200</td>
<td>Yes/Yes/Yes</td>
</tr>
<tr>
<td>CMU</td>
<td>150</td>
<td>Yes/Yes/No</td>
</tr>
<tr>
<td>UNC</td>
<td>150</td>
<td>No/No/Yes</td>
</tr>
</tbody>
</table>

Synchronization Evaluation

Video demo

Results for CMU dataset

Frontal view

Real video

Profile view

Depth map errors

Depth map from [1]

Depth map with our method

Groundtruth

Dense reconstruction visualization

Motivation
• Goal: Reconstruct a 3D image from a set of 2D X-ray projections.
• Novel scanner devices with spatially and temporally overlapping rays yield a new type of nonlinear ray constraints.
• Applications: medical imaging, industrial inspection, airport security, . . .

Contributions
• A new image reconstruction problem from X-ray measurements with overlap.
• A proof of partial convexity.
• A new optimization method based on forward-backward splitting.
• Experimental validation with real data.

Geometry of rays
Traditional CT: a single source  X-ray emitter array: multiple sources

Multiple (partially) simultaneously emitting sources lead to measurements with overlap:

Sequential scan: no overlap  Overlapping rays: several rays can reach the same detector at the same time

Acknowledgements
We thank Adaptil Ltd for providing the X-ray measurements used in the experiments. This work was supported by Adaptil Ltd and EPSRC EP/K003769/1.

A new type of ray constraints
Sequential measurements yield linear constraints:

\[ Ax = b \quad \text{with} \quad A_{ij} = -\xi_{ij} \]

Overlapping rays yield nonlinear constraints:

\[ \sum_{k=1}^{p} \exp \left( \sum_{i=1}^{n} -\xi_{ijk} x_i \right) = b_j, \quad \forall j = 1, \ldots, m \]

Sparse vectors of intersection lengths:

\[ r_{jk} = (-\xi_{jk,1}, \ldots, -\xi_{jk,n}) \in \mathbb{R}^n, \quad \forall j, k \]

A new image reconstruction problem
Lagrangian formulation with sparsity prior:

\[ \min_{x} \lambda \| x \|_1 + \frac{1}{2} \sum_{j=1}^{m} \left( \sum_{k=1}^{p} \exp \left( \sum_{i=1}^{n} -\xi_{ijk} x_i \right) - b_j \right)^2 \]

Possible choices for the regularization term:
• \( \| x \|_1 \): sparsity of densities
• \( \| x \|_p \): sparsity of gradients
• \( \| \Phi x \|_1 \): sparsity of wavelet frequencies

\[ \Phi \approx \exp \begin{pmatrix} -\xi_{11} & -\xi_{12} & \cdots & -\xi_{1n} \\ -\xi_{21} & -\xi_{22} & \cdots & -\xi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\xi_{m1} & -\xi_{m2} & \cdots & -\xi_{mn} \end{pmatrix} \]

Theorem:
The optimization problem is partially convex for \( x \leq \bar{x} \) and \( \nabla g \) is Lipschitz continuous with Lipschitz constant \( L = 2np^T \xi_{\max} \).

Forward-backward splitting optimization
Input:
\[ b \in \mathbb{R}^m : \text{measurements} \]
\[ r_{jk} \in \mathbb{R}^n : \text{intersection lengths} \]
\[ c \in (0, 1) : \text{line search control parameter} \]

Initialize \( x^0 = 0 \)
Iterate until convergence:

1. Compute search direction:
\[ \nabla g = \sum_{j=1}^{m} \left( \sum_{k=1}^{p} r_{jk}^T x_i - b_j \right) \left( \sum_{k=1}^{p} r_{jk} x_i \right) \]

2. Backtracking line search:
\[ \alpha = 1/L \]
\[ x^{n+1} = \text{prox}_{\lambda f}(x^n - \alpha \nabla g) \]
while \( \pi(x^{n+1}) < b \):
\[ \alpha = \alpha c \]
\[ x^{n+1} = \text{prox}_{\lambda f}(x^n - \alpha \nabla g) \]

3. Update \( x \):
\[ x^{n+1} = \text{prox}_{\lambda f}(x^n - \alpha \nabla g) \]

with \( \text{prox}_{\lambda f}(x) = \arg \min_y \left\{ f(y) + \frac{1}{2\lambda} \| y - x \|^2 \right\} \)

Reconstruction with ground truth data
Reconstruction of a cube from simulated data and reconstructed surface for \( p = 2 \):

Reconstruction from real-world measurements
DeeperCut is a deep, stronger, and faster multi-person pose estimation model that improves upon DeepCut, a previously developed method. It introduces several contributions that enhance its performance:

- **Deeper:** deeper architectures based on Residual Networks [3]
- **Stronger:** novel image-conditioned pairwise terms
- **Faster:** dramatic speed-ups due to strong pairwise and incremental optimization

**Contributions**

- A deeper, stronger and faster multi-person model
- + “deeper”: strong part detectors based on ResNet [3]
- + “stronger”: novel image-conditioned pairwise terms
- + “faster”: dramatic speed-ups due to strong pairwise and incremental optimization
- NEW: heuristic solver for real-time inference

**Unary Terms**

- deeper architectures based on Residual Networks [3]
- dilation and de-convolution reduce stride to 8 px
- joint reasoning at finest level of details
- weak pairwise based on geometry only
- infeasible run-time: takes hours to complete

**Pairwise Terms**

- image conditioned pairwise using CNN regression
- use regressed offsets and angles as features to train logistic regression to output pairwise probability

**Single Person Results**

<table>
<thead>
<tr>
<th>Setting</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Wri</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>PCKh</th>
<th>87.5%</th>
<th>85.2%</th>
<th>91.5%</th>
<th>89.9%</th>
<th>87.2%</th>
<th>90.1%</th>
<th>86.7%</th>
<th>90.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepCut [5]</td>
<td>97.4</td>
<td>92.5</td>
<td>88.1</td>
<td>85.9</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
</tr>
<tr>
<td>DeepCut (unary)</td>
<td>97.4</td>
<td>91.0</td>
<td>83.8</td>
<td>85.9</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
<td>87.1</td>
</tr>
</tbody>
</table>

**Multi-stage optimization**

- speed-up inference via incremental optimization
  1. solve for head and shoulder locations
  2. add elbows/wrists to stage 1 solution, re-optimize
  3. add rest of body parts to stage 2 solution, re-optimize

**Qualitative Multi-Person Results**

- **Successful cases**
- **Failure cases**

**References**


A New Idea for Representing Curves and Surfaces

Deformable registration techniques applied to point-sets often do not reflect the connectedness of the underlying shape. With the complex wave representation,

$$\psi(x) = \sum_{(m,n) \in S} \exp\left(-\frac{|x-m|^2}{2\sigma^2} + \frac{\nu \cdot (x-m)}{\lambda}\right),$$

we can add this feature (respecting the connectedness of a shape) to a registration pipeline that uses only normal- and point-based distance evaluation.

Figure 1: The surface is recovered using marching cubes on \(\psi_S\). Then it is painted with relative intensities from \(|\psi_S|^2\).

Properties of the Complex Wave Representation

Let \(\psi\) be the isotropic Gaussian:

- The representation distinguishes oriented point-sets:
  \(\psi\) is injective from \((\mathbb{R}^3 \times S^{n-1})^l\) into \(L^2(\mathbb{R}^3)\).
- It provides continuity of the normal field:\n  \(\hat{\nu}\) is defined a.e. with isolated (possible) singularities near Voronoi boundaries.
- It provides approximately linear composition:
  Given oriented point-sets \(S,T\), \(\psi_{S,T} \approx \psi_S \cdot \psi_T\).
- Given two oriented points satisfying certain basic conditions, the phase is continuous through a connected zero level-set.
- Near connections: \(|\psi|^2 > \sum_{m \in S} |T_m\psi|^2\).
- There are several closed-form distances between oriented point-sets induced by \(\psi\).

Setup for RDM

Deformation Level

Error \(10^{-2}\)

Occlusion Rate \(10^{-2}\)

Directions for Further Work

- Study different metrics for (2). anisotropic kernels.
- Organize both the template and the target.
- Alternate motion models.
- Improve the optimization strategy.
- Implicit representations of shape handle connection changes easily. Is there a spline or deformation model suited to the problem of registering different topologies?

Registration Results

We compared the registration performance of RDM against other point-based (and field-based) algorithms on the Gatsbain dataset.

Figure 3: Precision/Recall curves for the CMU-House dataset. ECCV paper also includes evaluation of RDM on 3D datasets.

Figure 4: Voronoi regions for a subset of correspondences obtained from RDM registration between two different subjects in similar poses.

Figure 5: A comparison of normal recovery approaches based on transferring the normal vector from a given registration vs. RDM. Simultaneously estimating the normal vector while solving the transforming problem leads to more reliable transformations and more accurate normal estimates.

Figure 6: RDM outperforms standard methods when considering multi-curve datasets. See the paper for a comparison of performance out-of-sample.

Figure 7: Examples of reconstructed curves using RDM.

Highlight

- \(\psi\) is a linear representation providing approximate signed distance in near field.
- Connects density-based shape approach with level-set approach, allowing to sample points or recover closed curves.
- A method for reconstructing while registering by combining density and geometry in 1 field.
- RDM shows that for certain tasks sparsely sampled shape information is enough to provide an improvement over point-based methods.

Code and Dataset Available

https://github.com/johncorring/RDM
- Web: http://www.cise.ufl.edu/~corring
- Email: johncorring@gmail.com
Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles
Mehdi Noroozi and Paolo Favaro
email: {noroozi,favaro}@inf.unibe.ch

Unsupervised learning of visual representations
- Visual representation is a configuration of parts (geometry) + parts appearance (texture)
- Can be learned by solving jigsaw puzzles as a pretext task
- Representation of a part focuses on differences rather than similarities

What does one need to learn to solve a jigsaw puzzle?

Solving Jigsaw Puzzles using the Context-Free Network (CFN)

ImageNet Results
- Classification results on ILSVRC 2012. Abscissa shows progressive locking of layers. All subsequent layers randomly initialized and retrained.

Pascal VOC2007 Results

Conclusions
- Solving jigsaw puzzles is a novel self-supervised learning task
- CFN architecture processes image tiles separately with little performance loss
- Feature transfer shows correlation to classification and detection tasks
- Task robust to ambiguities and easy to train
- Fast training time: 2.5 days
- Achieves state-of-the-art unsupervised representation

References

COCO Attributes
Attributes for People, Animals, and Objects
Genevieve Patterson  James Hays

COCO Attribute Dataset Statistics:
- 84,000 images
- 180,000 unique objects
- 196 attributes
- 29 object categories
- 3.5 Million objection-attribute pairs

Efficient Labeling Algorithm (ELA)
Exploiting Attribute Correlations

We don’t have to annotate all the attributes!

Person
male, sporty, participating, sitting, moving, what’s next?

Food
Carrot
healthy, tasty/delicious, fresh, what’s next?

Does the ELA work for every Attribute?

Comparing Classification:
1 vs. Rest and Multi-attribute estimation

Attribute labels are sparse. The average object has only 9 attributes (out of 196). Given an object’s category and known attributes, other attributes are more or less likely to be present.
Motivation

- Global motion compensation (GMC) removes intentional (due to camera pan/tilt/zoom) and unwanted (due to hand shaking) camera motion.
- Existing GMC algorithms sequentially register consecutive frames.
- Result is affected by temporal drift
- Failure of a pair of frames affects all upcoming frames.
- We propose a temporally robust GMC via joint alignment, in contrast to the sequential pipeline:
  - Drift problem is avoided
  - Abundance of information improves robustness

Formulation

- **Goal**: iteratively update the transformations applied to each frame so that the links connecting frames are parallel to \( t \)-axis and background regions are registered consistently.

- **Linearized via first-order Taylor expansion**:
  \[
  \Delta x(p) = \frac{\partial H(p)}{\partial p} \Delta p
  \]

- **Setting derivative to zero**:
  \[
  \Delta p = H^{-1}(p) \Delta x(p)
  \]

- **For the case of homography transformation**:
  \[
  v \equiv \frac{\begin{bmatrix} 1 & \frac{\partial h}{\partial x} \Delta x & \frac{\partial h}{\partial y} \Delta y \end{bmatrix} \begin{bmatrix} x \ y \ 1 \end{bmatrix}^T}{1 - x^2}
  \]

- **Alignment errors collected in the objective function**: \( \Delta x(p) = \Delta x^{(0)}(p) - \Delta x^{(1)}(p) \)

- Two factors considered in setting the weights:
  - Keypoints detected at larger scale are more likely to be from background matches. \( \Delta x^{(0)}(p) \) is set proportional to the scale: \( M_x(p) \).
  - For each frame, the links may be made either to:
    - All the previous frames (backward scheme)
    - Both the previous and upcoming frames (backward-forward scheme) with different weights.

- **Alignment of non-keyframes**:
  - Each non-keyframe aligned independently with 2 surrounding keyframes

- **Evaluation Metric**: Background region error (BRE):
  \[
  BRE(i) = \frac{1}{|M_{d_{\ell}}|} \sum_{p \in M_{d_{\ell}}} \left( I(p) - I'(p) \right)
  \]

- **Comparison of GMC algorithms on quantitative dataset (55 videos)**

- **Comparison of GMC algorithms on qualitative dataset (200 videos)**

The source code is available at: http://www.cvlab.cse.msu.edu/project-trgmc.html
(A) GOAL & CONTRIBUTIONS

Recently, several methods to understand CNNs through visualization have been proposed:
1. DeConvNet visualizes patterns selected by neurons [5].
2. Class saliency visualizes the “network attention” pattern [3]. However, both are heuristic and their meaning remains unclear.

Our goal is to unify, compare, and understand such techniques. We do so by:
1. Introducing a generalized construction for reversed architectures.
2. Exploring in detail three variants: DeConvNet, SaliNet and the hybrid DeSaliNet.
3. Identifying limitations of these networks for the purpose of CNN visualization.

(B) Reversing CNN Architectures

Reversing a architecture layer by layer

Reversing layers

Back-propagation defines a natural reverse of each layer. For layer \( f, x \rightarrow y \) then its BP-reversed becomes

\[ y' = \phi(y) \leq \phi(x) > x', \phi(x) > x' \]

where \( y' \) is the layer input for the reversed layer.

However, other definitions are also commonly used.

We consider variations used:

Deconvolutional Networks (DeConvNet) – Zeiler et al.
Class Saliency (SaliNet) – Simonyan et al.
Improved DeConvNets (DeSaliNet) – Springenberg et al.
Backpropagation – Rumelhart et al.
Semantic Segmentation U-Net – Noh et al.

(C) Reversing Layers: Max Pooling and ReLU

Max Pooling

ReLU

Pooling

ReLU

Rectification Mask \( I \)

ReLU

ReLU

ReLU

No Operation

(D) Analysis of Reversed Architectures

Two types of reversed pooling and four types of reversed ReLU layers. Result images are shown below.

(E) Lack of Neuron Selectivity

We change the neuron selector and view the result image

DeConvNet

SaliNet

DeSaliNet

Changing the selected neuron does not significantly change the output.

Suggests that these reversed networks are not suited for neuron visualization.

The auxiliary information dominates the output.

(F) Interpretation: Fourier Phase Information

Input Image

Fourier Reconstruction

Fourier Reconstruction Random Magnitude

DeSaliNet Random Neural Input

DeConvNet Positive Random Input

Auxiliary information is like the phase information in a Fourier transform.

Randomized magnitudes with ground truth phase yields an edge image similar to the results obtained from a reversed CNN using ReLU backward.

(G) Foreground Object Selectivity

- Foreground background differentiation is perhaps implicit in the CNN hidden layer activations.
- Extract this and project it into the image using a “Reversed CNN.”
- \( \text{VGG-VD-Pool5}_3 \) – \( \text{VGG-VD-FC8} \)

(H) Weakly Supervised Foreground Object Segmentation

- Use the output of a “Backward CNN” to seed a grab cut segmentation. Segment the foreground object!
- We compare against the weakly supervised baseline of Guillaumin et al. IJCV 2014

References


Acknowledgements: BP for Aravindh Mahendran, ERC SIG IDIU for Andrea Vedaldi.
Motivation
- Image priors play a key role in low-level vision tasks
- Different priors capture different geometric properties
- Our visualization deforms images to best conform to a given prior

Optimization Problem
\[
\arg\min_{x,T} - \log p(x) + \lambda \Phi(T) + \frac{1}{2\sigma^2} \|T(y) - x\|^2
\]
For any input image \( y \), we seek for \( x \approx T\{y\} \), for some smooth deformation \( T \), such that \( \log p(x) \) is maximal.

Algorithm
Alternating Minimization:
- \( \mathcal{T} \)-step: \( \arg\min_x \|T(y) - x\|^2 - \log p(x) \)
- \( \mathcal{O} \)-step: \( \arg\min_T \|T(y) - x\|^2 + 2\sigma^2 \cdot \Phi(T) \)

Results

Visualization strength
The algorithm converges when:
\[
T = \text{OpticalFlow}(y, \text{Denoise}(T\{y\}));
\]
The geometry of \( T\{y\} = y^{\text{GEM}} \) is not altered by the Denoiser.
INTRODUCTION

- **Overview:** Estimating robust and accurate optical flow.
- **Input:**
  - Monocular temporal image pair
- **Output:**
  - Dense vector field linking each pixel in first image with corresponding pixel in second image
- **Our Contributions:**
  - A novel optical flow technique for autonomous driving that exploits instance level semantic information, assumes rigid body motion, and enforces epipolar constraints for each car
  - State of the art result on KITTI Flow 2015 benchmark
  - A novel siamese deep neural network pipeline for corresponding pixel matching and uncertainty estimation

DEEP MATCHING

- **Training:** We train our matching CNN with a small left image patch and a bigger right image strip. The right image strip is cropped vertically and horizontally.
- **Cost Function:** We apply softmax over all disparities within search window. We use cross-entropy with reduced penalty for small offset error.
  \[
  \min_w \sum_{i=1}^{N} \sum_{s_i} p_i^{GT}(s_i) \log p_i(s_i, w)
  \]
  \[
  p_i^{GT}(s_i) = \begin{cases} 
  \lambda_1 & \text{if } s_i = s_i^{GT} \\
  \lambda_2 & \text{if } |s_i - s_i^{GT}| = 1 \\
  \lambda_3 & \text{if } |s_i - s_i^{GT}| = 2 \\
  0 & \text{otherwise}
  \end{cases}
  \]
- **Inference:** We apply the network to entire image with 400 \(\times\) 200 search window. We save only top-K matches to save memory.
- **Confidence map:** We perform cost aggregation on top-K volume. The matching score provides a confidence measure. We use thresholding to generate the initial flow image.

EPIPOLAR FLOW ESTIMATION

- We use instance segmentation to separate possibly moving cars from background.
- We estimate flow for each object and background independently by exploiting epipolar constraints (i.e., each object moves rigidly in 3D).
- We use the 8-pt algorithm to estimate fundamental matrices with matching network output.
- We separate flow into linearized flow from relative rotation + flow from relative translation which is directly toward or away from epipole of second image. This is a 1D search problem for matches.
- We calculate the 1D disparities with Semi-Global matching.

\[
E(d) = \sum_{p_k} C^p_k(p_k, d_{p_k}) + \sum_{p_k, p_k' \in \mathcal{N}} S(d_{p_k}, d_{p_k'})
\]

\[
S(d_{p_k}, d_{p_k'}) = \begin{cases} 
\lambda_1 & \text{if } |d_{p_k} - d_{p_k'}| = 1 \\
\lambda_2 & \text{if } |d_{p_k} - d_{p_k'}| > 1 \\
0 & \text{otherwise}
\end{cases}
\]

- We smooth the resulting output with EpicFlow for foreground objects. We use linearized epipolar interpolation and slanted plane fitting for the background.

EXPERIMENTS

- **Quantitative results:** KITTI Flow 2015 Test set results (error rate comparisons to other monocular approaches)

<table>
<thead>
<tr>
<th>Method</th>
<th>F1-flg</th>
<th>F1-flg</th>
<th>F1-all</th>
<th>F1-flg</th>
<th>F1-flg</th>
<th>F1-all</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td>30.49%</td>
<td>30.59%</td>
<td>34.13%</td>
<td>39.90%</td>
<td>83.39%</td>
<td>42.18%</td>
</tr>
<tr>
<td>DeepFlow</td>
<td>16.47%</td>
<td>31.25%</td>
<td>19.15%</td>
<td>27.96%</td>
<td>35.28%</td>
<td>29.18%</td>
</tr>
<tr>
<td>EpicFlow</td>
<td>15.00%</td>
<td>29.39%</td>
<td>17.61%</td>
<td>28.81%</td>
<td>33.56%</td>
<td>27.10%</td>
</tr>
<tr>
<td>MotionSLIC</td>
<td>6.19%</td>
<td>64.82%</td>
<td>16.83%</td>
<td>14.86%</td>
<td>66.21%</td>
<td>23.40%</td>
</tr>
<tr>
<td>DiscreteFlow</td>
<td>9.96%</td>
<td>22.17%</td>
<td>12.18%</td>
<td>21.53%</td>
<td>26.68%</td>
<td>22.38%</td>
</tr>
<tr>
<td>SOF</td>
<td>8.11%</td>
<td>23.29%</td>
<td>10.86%</td>
<td>14.63%</td>
<td>27.25%</td>
<td>16.91%</td>
</tr>
<tr>
<td>Ours</td>
<td>3.75%</td>
<td>22.28%</td>
<td>8.79%</td>
<td>8.61%</td>
<td>26.69%</td>
<td>11.62%</td>
</tr>
</tbody>
</table>

- **Qualitative results:** within each group, top to bottom: one input image, matching network output, external instance segmentations (x2), flow output, error image.
Introduction

Ultrasound (US) is widely used in medical diagnostics and during therapy as a low-cost, flexible, and real-time imaging technique. In fetal medicine, US is both used for non-invasive diagnostics and to guide surgical interventions.

pose tracking and calibration of a 3D US probe enhances its applications to Computer Assisted Intervention (CAI) in the following aspects:

- Building large and detailed 3D models from several 3D US frames.
- Enabling freehand 4D US, i.e. registering both a 3D volume and its temporal evolution in a single coordinate system while the probe is being freely moved.
- Registration between 3D US data and other instruments, such as biopsy needles.

Calibrating a tracked US probe is done by scanning a calibration target with known shape. When using a curvilinear US probe, this is a similarity registration problem (also known as Pose-and-Scale problem), and it involves estimating both a rigid transformation and a scale factor that converts US scan units to metric coordinates.

In this paper we use a tracked needle as the calibration target (a linear object) and the calibration problem becomes the similarity registration between two sets of 3D lines (using a 2D US probe), or the similarity registration between co-planar points and 3D lines (using a 2D US probe).

Formulation

SCANNING A TRACKED NEEDLE

A needle is measured as a point in a 2D US, and as a line segment in a 3D US.

CONVERTING TO (POINT ↔ PLANES) CORRESPONDENCES

Both problems are converted to the similarity registration between 3D points and 3D planes by defining each line li as two planes, and each line bi as two points X_i, X_i'.

GENERAL SOLUTION

The general problem has 13 unknown linear parameters and it can be linearly solved from 6 point-line correspondences or 3 line-line correspondences. However, A has 5 independent quadratic constraints, and thus it can be minimally solved from 4 point-line correspondences or 2 line-line correspondences.

SIMPLIFIED SOLUTION (CO-PLANAR POINT ↔ LINE)

When the point correspondences are co-planar (2D US), even though the general solution works, the problem can be simplified to 10 unknown linear parameters and it can be linearly solved from 5 point-line correspondences.

Results

SYNTHETIC DATA

Our method is tested in a virtual environment where a fixed 2D/3D US probe scans a tracked needle at 3D different random positions. The US probe has a depth range of 107mm, angle range [−50, 50], a scale factor (s = 0.24 mm/pix). Both the US scans and the needle tracking measurements are injected with Gaussian noise (σ = 2 pix, α = 1mm respectively). The results are compared against groundtruth values of rotation, translation, and scale factor.

REAL DATA

Our algorithms are tested using a GE Voluson E10 machine with a c5-1 probe (3D US). The scanning depth is set to 107mm, and both 3D US and 2D US data are obtained with the same equipment and settings. Needle tracking is achieved with Optiptrack V120 Trio. The US measurements are performed in a container filled with water at room temperature.

After 20 calibration trials, we scan a 3D point target and measure its projection reconstruction accuracy (PRA), i.e., the difference in mm between the 3D point location measured using the needle tip and its projection from the US scan. We performed 10 acquisitions of the 3D point target. Each distribution contains 200 error measurements.

Conclusions

Linear and minimal solutions are tested to calibrate a US probe using a tracked needle with both 3D and 2D data. This is useful in medical imaging to guide a biopsy needle in US based interventions. The method can be easily extended to additional US calibration problems using other types of calibration targets, e.g. scanning single plane target leads to the similarity registration between co-planar lines and 3D planes (2D US) or between two sets of 3D planes (3D US). In other computer vision domains this algorithm can potentially be used as an extension of the pose-and-scale problem to the alignment of line-based and/or plane-based SFM sequences.

Acknowledgements

This work was supported through an Innovative Engineering for Health award by the Wellcome Trust [WT110195]; Engineering and Physical Sciences Research Council (EPSRC) [NS/A000027/1]. Danail Stoyanov receives funding from the EPSRC [EP/P013220/1, EP/N022750/1], the EU-FP7 project CASCADE(FP7-ICT-2913-601021) and the EU-Horizon2020 project EndoVESPA (H2020-ICT-2015-688592).

References

[27] Bartoli, Sturm The 3d line motion matrix and alignment of reconstructions. CVPR 2001
[28] Ramanujan et. al. P2n: A minimal solution for registration line of 3d points to 3d planes. ECVIC 2010
Figure: Approximate recovery of a toy example on a cat mesh (N = 290). a) Band-limited random signal in [0, 1) with noise on the cat mesh. b) Expected sampling point of the given method with k = 40. c) Recovered signal using our method with k = 50. d) Recovered signal using our method with k = 100. Note that our recovery is estimating only 50 parameters instead of N = 290.

Signal Recovery of Band-limited Signals: Algorithm
- Select \( \Omega \) using \( \mathbf{p}_s \).
- Define a projection operator \( M_{\Omega, j} \) (i.e., a sampling matrix) yielding \( M_{\Omega}^{j} = y \) as
  \[
  M_{\Omega, j} = \begin{cases} \mathbf{I} & \text{if } j = \omega_j \\ 0 & \text{o.w.} \end{cases}
  \]
- Let \( \hat{g}_{\gamma}(\omega) = \sum_{k=0}^{N} \hat{g}(\omega_k) \cdot e^{i \omega_k} \) and \( \hat{p}_s \) be the first k coefficients.
- We define and solve a convex optimization problem:
  \[
  \min_{\hat{g}} \left\{ \| M_{\Omega} \hat{g} - y \|_2^2 + \gamma \| \hat{g} \|_1 \right\}
  \]
  where \( \gamma \) is estimated as \( \hat{g} - V \hat{g}_s \).
- An analytic solution \( \hat{g}_s \) must satisfy the condition
  \[
  \left( V^T M_{\Omega} V + \gamma I \right) \hat{g}_s = \hat{g} - V \hat{g}_s
  \]
- Using the optimal \( \hat{g}_s \), we recover a band-limited signal as \( g^* = - V \hat{g}_s \).

Brain Imaging Modalities for Experiments
- Pittsburgh compound B (PIB) PET scan: captures protein levels in the brain
- Tractography from Diffusion Tensor Image (DTI): provides structural brain connectivity

Predicting FA using Covariates in HCP
- The Human Connectome Project (HCP) Dataset. (N = 487)
  - Inexpensive: 27 covariates (cognition, demographics, education, etc.) related to AD.
  - Expensive: 17 Fractional Anisotropy (FA) measurements of major fiber bundles.

Goal. Recover FA measurements on all 487 participants using a) covariates from all 487 participants and b) FA measurements from only m = 487 participants.
Result. Distribution of t-form errors with 100 runs: 1) Ours (lower errors), 2) Rao et al. (NIPS2015) and 3) Puy et al. (Applied and Computational Harmonic Analysis 2016).

Spheres representing FA prediction errors (t-form) at each fiber bundle in the HCP study. Our results (left) show smaller errors (error) than Puy et al. (middle) and Rao et al. (right).

Acknowledgment
Research supported by NIH grants AG040396, NSF CAREER award 1252725, UW ADRC AG033514, UW ICTR 1UL1RR025011, UW CPRP AT117524, UW CBTRUST U54MH087523-14 and Waisman Core Grant P50 HD015925-45.

European Conference on Computer Vision (ECCV) 2016
1. Motivation

1. Skin disease is one of the most common illnesses in human daily life. It affects between 30% and 70% of individuals. However, diagnosis of skin diseases by observing is a very difficult job, where an intelligent system can be helpful.

2. Most previous works on recognition of skin disease are restricted to dermoscopic images, and the research about clinical images is scant. Moreover, all the existing clinical skin disease datasets are small scale and not available, so the lack of datasets is a barrier to this domain.

3. In contrast to scene classification or object classification, skin disease has its own characteristics. But, there is no research to discuss and explore how to catch the characteristics of clinical skin disease and represent them through computer vision technology.

2. Contribution

“SD-198” is currently the largest and available clinical skin disease dataset, which has been collected and publicly released in our work.

1. It contains 198 skin diseases and 6,584 clinical images.

2. All images are collected from the real world, generated and uploaded by patients and doctors.

3. It cover a lot of situations for patients such as age (child, adult, old), sex, disease site (hand, feet, head, nails), color of skin and different periods of lesion (early, middle, late) and with variance in color, exposure, illumination and level of details.

To spur further related research, we select a subset SD-128 (the number of images>20 ) from SD-198.

The performance of skin disease classification using deep features as well as hand-engineered features has been evaluated in our work.

3. Dataset

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>128</td>
<td>198</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Images</td>
<td>2309</td>
<td>5619</td>
<td>6584</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>2012</td>
<td>2013</td>
<td>2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Available?</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

4. Creation, Examples and Distribution of SD-128 and SD-198

5. Experimental Results and Analysis

Results with hand-engineered features

<table>
<thead>
<tr>
<th>#</th>
<th>Feature</th>
<th>Feature Dimension</th>
<th>SD-198[%]</th>
<th>SD-128[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SIFT</td>
<td>21,000</td>
<td>25.85</td>
<td>29.40</td>
</tr>
<tr>
<td>2</td>
<td>HOG</td>
<td>12,400</td>
<td>12.78</td>
<td>14.17</td>
</tr>
<tr>
<td>3</td>
<td>LBP</td>
<td>21,200</td>
<td>15.46</td>
<td>17.09</td>
</tr>
<tr>
<td>4</td>
<td>Color Histogram</td>
<td>768</td>
<td>4.19</td>
<td>5.59</td>
</tr>
<tr>
<td>5</td>
<td>Color Names</td>
<td>21,000</td>
<td>20.20</td>
<td>20.32</td>
</tr>
<tr>
<td>6</td>
<td>Gist</td>
<td>512</td>
<td>16.49</td>
<td>17.52</td>
</tr>
<tr>
<td>7</td>
<td>Gabor</td>
<td>4,000</td>
<td>10.14</td>
<td>11.37</td>
</tr>
</tbody>
</table>

Results with deep features

<table>
<thead>
<tr>
<th>Deep network</th>
<th>SD-198[%]</th>
<th>SD-128[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CaffeNet</td>
<td>42.31</td>
<td>42.83</td>
</tr>
<tr>
<td>CaffeNet+ft</td>
<td>46.69</td>
<td>47.38</td>
</tr>
<tr>
<td>VGG</td>
<td>37.91</td>
<td>39.27</td>
</tr>
<tr>
<td>VGG+ft</td>
<td>50.27</td>
<td>52.15</td>
</tr>
</tbody>
</table>

6. Reference


Deep Learning 3D Shape Surfaces using Geometry Images

Motivation
Learning 3D shape volume surface using voxel geometry image representation

Why Geometry Images?
- Representation is 2D while encoding pertinent shape information, reducing memory and computational complexity.
- Can be flexibly encoded with intrinsic or extrinsic signatures suitable for non-rigid and rigid shape analysis using convolutional neural networks (CNNs).

Contributions
1. Established relevance of authalic spherical parametrization for creating geometry images used subsequently in CNN.
2. Robust authalic parametrization of arbitrary shapes using area restoring diffeomorphic flow and barycentric mapping.
3. Creation of geometry images (a) with appropriate shape feature for rigid/non-rigid shape analysis, (b) which are robust to cut and amenable to learn using CNNs.

Data Preprocessing
1. Voxelize shape to follow Euler characteristic: $2 - 2m = |V| - |E| + |F|$
2. Convert to genus-0

Cut & Data Augmentation
Cuts serve to augment data and rotated spherical parametrizations create images which are projected from different viewing directions.

Shape Classification and Retrieval

<table>
<thead>
<tr>
<th>Model</th>
<th>McGill1 Classify</th>
<th>McGill1 Retrieve</th>
<th>SHREC1 Classify</th>
<th>SHREC1 Retrieve</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShapeNet</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
</tr>
<tr>
<td>Conformal</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
</tr>
<tr>
<td>SPHARM</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
</tr>
<tr>
<td>Ours</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
</tr>
</tbody>
</table>

Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>McGill1 Classify</th>
<th>McGill1 Retrieve</th>
<th>SHREC1 Classify</th>
<th>SHREC1 Retrieve</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShapeNet</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
</tr>
<tr>
<td>Conformal</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
</tr>
<tr>
<td>SPHARM</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
</tr>
<tr>
<td>Ours</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>McGill1 Classify</th>
<th>McGill1 Retrieve</th>
<th>SHREC1 Classify</th>
<th>SHREC1 Retrieve</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShapeNet</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
</tr>
<tr>
<td>Conformal</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
</tr>
<tr>
<td>SPHARM</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
</tr>
<tr>
<td>Ours</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
</tr>
</tbody>
</table>

Why Spherical Authalic Parametrization?
- Representation is 2D while encoding pertinent shape information, reducing memory and computational complexity.
- Can be flexibly encoded with intrinsic or extrinsic signatures suitable for non-rigid and rigid shape analysis using convolutional neural networks (CNNs).

How do we do parameterize?
1. Solve a Poisson equation to get area distortion field
2. Convert to genus-0
3. Barycentric map area restoring field on the shape onto a sphere. Does not require retriangulation or recalculating Laplacian pseudoinverse, hence, efficient.

Creating Geometry Images for CNN Learning

<table>
<thead>
<tr>
<th>Model</th>
<th>McGill1 Classify</th>
<th>McGill1 Retrieve</th>
<th>SHREC1 Classify</th>
<th>SHREC1 Retrieve</th>
</tr>
</thead>
<tbody>
<tr>
<td>ShapeNet</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
<td>65 0.29 57.2 0.28</td>
<td>52.7 0.1 48.4 0.13</td>
</tr>
<tr>
<td>Conformal</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
<td>55 0.36 58.0 0.45</td>
<td>58.0 0.45 58.0 0.45</td>
</tr>
<tr>
<td>SPHARM</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
<td>62 0.35 82.5 0.58</td>
<td>82.5 0.58 82.5 0.58</td>
</tr>
<tr>
<td>Ours</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
<td>83 0.75 92.5 0.72</td>
<td>92.5 0.72 92.5 0.72</td>
</tr>
</tbody>
</table>

Seamless padding informs CNN about warped geometry and implicitly makes learning robust to cut.

Shape Analysis Pipeline

1. Established relevance of authalic spherical parametrization for creating geometry images used subsequently in CNN.
2. Robust authalic parametrization of arbitrary shapes using area restoring diffeomorphic flow and barycentric mapping.
3. Creation of geometry images (a) with appropriate shape feature for rigid/non-rigid shape analysis, (b) which are robust to cut and amenable to learn using CNNs.

Why Spherical Parametrization? Why Authalic Parametrization?
1) Under constraint of resolution in geometry images, authalic parametrization preserves more shape information.
2) Compatible with notion of convolving a shape patch with fixed filter sizes in a CNN.

How do we do parameterize?
- Barycentric map area restoring field on the shape onto a sphere. Does not require retriangulation or recalculating Laplacian pseudoinverse, hence, efficient.

Awan Sinha
sinha12@purdue.edu

Jing Bai
bai58@purdue.edu

Karthik Ramani
ramani@purdue.edu

C Design Lab
https://engineering.purdue.edu/cdesign/
Deep Image Retrieval: Learning Global Representations for Image Search
Albert Gordo, Jon Almazán, Jérôme Revaud, and Diane Larlus
Computer Vision Group, Xerox Research Centre Europe

Motivation
Task: instance-level image retrieval

Underwhelming results of deep learning methods

Problems:
- Architecture
- Training data
- Training procedure

R-MAC as a differentiable network

Original R-MAC:
- Feature extraction + ad-hoc aggregation and normalization

Observations:
- Aggregation and normalization can be integrated in the network
- The resulting descriptor is fully differentiable

Improving the R-MAC representation

Beyond fixed regions
- Rigid grid is suboptimal: many regions cover only background
- Interest region prediction with region proposal network

At test time: everything extracted in a single forward pass

Multi-resolution
- Features extracted at different scales and aggregated

Data cleaning

Public dataset of landmark images:
~200K images
~600 different landmarks
Dataset is very noisy

Cleaning:
1. Keypoint matching of all image pairs in each landmark
2. Construct a graph with matching images
3. Bounding box prediction using keypoints + diffusion process

At the end:
~40k spatially verified images
+ Approximate bounding box annotations

Training

Learning to rank
- Triplet ranking loss:
  \[ \frac{1}{2} \max(0, m + \|q - d^+\|^2 - \|q - d^-\|^2) \]
- Three-stream siamese network

Experiments

Evaluation on five datasets:
- Oxford 5k
- Oxford 105k
- Paris 6k
- Paris 106k
- Holidays

Baseline:
Trained with classification loss
Trained with ranking loss
Trained with ranking loss + proposals
Trained with ranking loss + multi-resolution
Trained with ranking loss + proposals + multi-resolution

Observations:
- Learning with clean data is critical
- Ranking loss > classification loss
- Proposals and Multi-resolution help, particularly with VGG16
- ResNet101 > VGG16

Comparison with state of the art

Short codes (PQ)

Long codes and matching-based methods

State-of-the-art results on the five datasets
Motivation

- Scene understanding benefits from rich 3D scene models with complete depth and semantics.

- Depth sensors produce imperfect measurements: sparse, containing large holes and subject to occlusion.

Contributions

We introduce a fully-automatic approach that
- jointly completes and hallucinates visible and hidden depth and semantics;
- encodes a piece-wise planar world assumption;
- efficiently minimizes a discrete-continuous energy function based on the Mumford-Shah functional [1].

Model Overview

Two-layer scene Model:

\[
\begin{aligned}
\text{Visible layer, } (u^v, s^v) & \quad \rightarrow \quad \text{Scene} \\
\text{Hidden layer, } (u^h, s^h) & \quad \rightarrow \quad \text{Mask, } m
\end{aligned}
\]

Notations:

- \( x \in \Omega \): pixel location.
- \( s^v(x) \in \mathbb{R}^{2} \): Visible semantic label (for \( L \) classes).
- \( y^v(x) \in \mathbb{R} \): Visible disparity value; plane representation \( y^v(x) = p(x)^T u^v(x) \), where \( p(x) = (x^2, 1)^T \).
- \( u^v(x) \in \mathbb{R}^3 \) and \( s^h(x) \in \mathbb{R}^{2} \): disparity value and semantic label of the hidden scene layer.
- \( m(x) \): binary mask indicating the foreground class. When \( m(x) = 1 \), there are neither disparity measurements nor semantic predictions for the hidden layer variables \( u^h(x) \) and \( s^h(x) \).

Model Formulation

Two-layer Scene Model for joint depth and semantics completion and hallucination:

\[
\begin{aligned}
\min_{u^v, s^v, m} & \quad E_v + E_{\alpha} + E_v + E_{\beta} + E_s \\
\text{visible layer} & \quad \text{hidden layer} \\
\text{s.t.} & \quad u^v(x) = u^v(x), s^v(x) = s^v(x) \quad \forall x \mid m(x) = 0 \\
& \quad \sum_{k} s^v_k(x) = 1, s^v_k(x) \geq 0, \quad \forall x, k \\
& \quad \sum_{k} s^h_k(x) = 1, s^h_k(x) \geq 0, \quad \forall x, k \\
& \quad m(x) \in \{0, 1\}, \quad \forall x
\end{aligned}
\]

Visible semantic and depth layer:

- \( E_d(u^v, s^v) \) encourages the model variables to be consistent with the disparity measurements and the estimated semantic probabilities.
- \( E_{\alpha}(u^v, s^v) \) encourages both \( u^v \) and \( s^v \) to be piecewise constant while having their discontinuities aligned.
- \( E_v(s^v, m) \) encourages the agreement between \( s^v \) and \( m \).

Hidden semantic and depth layer:

- \( E_{\alpha}(u^h, s^h, m) \) encourages \( u^h \) and \( s^h \) to be piecewise constant and have aligned discontinuities.
- \( E_h(s^h, m) \) encourages \( s^h \) to be consistent with the statistics of average disparity map for each background class

Layer constraints:

- Agreement between the two layers.

\[ u^v(x) = u^h(x), s^v(x) = s^h(x), \forall x \mid m(x) = 0 \]

Piece-wise constant regularization [1]:

- Non-convex Mumford-Shah functional

\[
\min_D(D(y) + R(Ay))
\]

- \( D(\cdot) \) denotes a data fidelity term;
- \( R(\cdot) \) is the regularization term encouraging piece-wise-smoothness.

A denotes a linear operator, which is an (oriented) gradient operator.

Solved by primal-dual method.

Model Optimization

Initialization:

- \( u^v \) from sparse visible depth and \( s^v \) using FCN-32s.
- occluded part of \( u^h \) and \( s^h \) as 0.

Alternating minimization:

- With \( m \) fixed, the energy decomposes into two sub-problems: one for the visible layer, and one for the hidden one.
  - Each one corresponds to a multimodal version of the Mumford-Shah functional.
- Update \( m \) (closed form).

Experimental Results

Baseline for depth:

- Visible layer: classical method [2], and more recent technique of [4].
- Hidden layer: Baseline-1 (semantic segmentation followed by [2] + [3]) and Baseline-2 (semantic segmentation followed by [4] + [3])

Quantitative results for visible, hidden depth and semantics:

<table>
<thead>
<tr>
<th>visible-acc</th>
<th>baseline-1</th>
<th>baseline-2</th>
<th>baseline-3</th>
<th>baseline-4</th>
<th>baseline-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>58.8%</td>
<td>58.0%</td>
<td>51.2%</td>
<td>45.8%</td>
<td>45.7%</td>
</tr>
<tr>
<td>[4]</td>
<td>53.4%</td>
<td>53.7%</td>
<td>48.3%</td>
<td>45.0%</td>
<td>47.5%</td>
</tr>
<tr>
<td>[2]</td>
<td>50.4%</td>
<td>52.7%</td>
<td>47.3%</td>
<td>44.6%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

Quantitative results for visible depth:

<table>
<thead>
<tr>
<th>visible-acc</th>
<th>baseline-1</th>
<th>baseline-2</th>
<th>baseline-3</th>
<th>baseline-4</th>
<th>baseline-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>58.8%</td>
<td>58.0%</td>
<td>51.2%</td>
<td>45.8%</td>
<td>45.7%</td>
</tr>
<tr>
<td>[4]</td>
<td>53.4%</td>
<td>53.7%</td>
<td>48.3%</td>
<td>45.0%</td>
<td>47.5%</td>
</tr>
<tr>
<td>[2]</td>
<td>50.4%</td>
<td>52.7%</td>
<td>47.3%</td>
<td>44.6%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

Quantitative results for hidden semantics:

- per class accuracy: 35.42% \( \pm \) per pixel accuracy: 50.08%

References:

Motivation

- **Goal**: Efficient video navigation
  - Automatic spatio-temporal localization of concepts
- **Problem**: Non scalability
  - Manual annotation
  - Explicit learning of million concepts
- **Old idea**: Millions of weakly labeled videos on internet
  - Localization information is missing.
  - But concept co-occurrence prior can be easily extracted.

Contribution

- Jointly learn person, objects & action model from weakly labeled videos.
- Propose novel Indian Buffet process model which incorporates location constraints and heterogeneous features in an integrated framework.
- Posterior inference of proposed model using mean-field approximation.
- Improved result for concept classification and localization across datasets.

Bayesian Model Learning

M: Total number of videos
\( \alpha_1, \alpha_2, \alpha, \alpha_1, \alpha_2 \): Hyperparameters
\( \mathbf{A}_i, \mathbf{A}, \sim \mathcal{N}(0, \mathbf{I}) \): Appearance model

For each video \( i \in 1, \ldots, M \),
\( x_i^j \sim \text{Stick}(\alpha) \): Video prior
\( \mathbf{L}_{vi} \in [0, 1] \): Weak input labels for kth concept

For each track \( j \in 1, \ldots, N_i \),
\( \mathbf{x}_i^j \sim \text{Beta}(\mathbf{z}_i^j, \mathbf{t}_i^j) \): Latent coefficient
\( \mathbf{a}_i^j \sim \mathcal{N}(\mathbf{w}_i^j, \mathbf{A}^+, \mathbf{I}) \): Feature vector

Learning requires computing posterior distribution of latent variables,
\( Y = \{ A, \pi^1, \mathbf{Z}^1 | i \in 1, \ldots, M \} \),
\[ q(Y) \propto \prod_{i=1}^{M} \prod_{j=1}^{N_i} \mathcal{N}(\mathbf{a}_i^j | \mathbf{w}_i^j, \mathbf{A}^+, \mathbf{I}) \]

New posterior is constrained by the location constraints prior,
\[ \sum_{j \in \mathbf{Z}_i} x_i^j \geq \sum_{j \in \mathbf{Z}_i} \mathbf{a}_i^j \geq 1 - \mathbf{z}_i^1 \]

Minimize KL divergence \( D(q(Y)||p(Y|X, \Theta)) \) under expectation constraint

Exact inference intractable. Tractable parametrized family of distribution is used,
\[ q(Y) = \prod_{i=1}^{M} \prod_{j=1}^{N_i} \mathcal{N}(\mathbf{a}_i^j | \mathbf{w}_i^j, \mathbf{A}^+, \mathbf{I}) \]

WSC-SIIBP Algorithm

- **Input**: Training and Test data feature vector and train video weak label.
  - Step 1: Initialize both old and new distribution hyperparameters.
  - Step 2: Update new hyper parameters \( \nu^j, \Phi, \Psi, \) for \( t \) iterations.
  - Step 3: Update \( \mathbf{a}_i^j, \mathbf{e}_i^j, \), \( \mathbf{r}_i^j \) Repeat Step 2 for \( t \) iterations.

Output: Posterior of Y. Test video labels are inferred from \( \nu^j \).

Experiments

Datasets

- Caba1abana movie
  - 19 persons, 3 actions
  - Segments 60 / 120 s
  - 1094 face tracks
- A2D
  - 7 objects, 9 actions
  - 7-10 sec long
  - 3782 YT videos

Face features

- Face detection
- Object extraction
- Dense root SIFT
- PCA
- Fisher vector

Action features

- Extrapolate face track bounding box
- Dense trajectories
- Fisher vector
- PCA
- Fisher vector

Weakly Supervised Learning of Heterogeneous Concepts in Videos

References


This work began when Sohil and Kuldeep were interns at Xerox Research Center India, Bangalore. Xerox Research Center India is sponsoring Sohil’s travel to ECCV.
Recurrence Instance Segmentation
B. Romera-Paredes, P. H. S. Torr
University of Oxford, UK

Instance Segmentation Problem

Instance segmentation is the problem of detecting and delineating each distinct object of interest appearing in an image. Most approaches proposed for instance level segmentation are based on a pipeline of modules whose learning process is carried out independently of each other.

Our approach

Humans count sequentially, using spatial memory in order to keep track of the accounted locations [4]. Driven by this insight, our purpose is to build a learning model capable of segmenting the instances of an object in an image sequentially, keeping the current state in an internal memory.

We rely on recurrent neural networks (RNNs), which exhibit both the ability to produce sequential output, and the ability to keep a state or memory along the sequence.

Our primary contribution is the development of an end-to-end approach for instance segmentation based on:

- RNNs containing convolutional layers.
- A principled loss function for instance segmentation.

ConvLSTM

We are interested in recurrent structures based on convolutions, in which the intermediate representations of the images preserve the spatial information.

Spatial inhibition module

Two outputs are produced by our model at each time:

1. A map that indicates which pixels compose the object that is segmented in the current iteration. Thus, the function learned for this stage has to be able to discriminate one, and only one instance, filtering out everything else.
2. The estimated probability that the current segmented candidate is an object.

Loss function

Our model predicts both a sequence of masks, \( \hat{Y} = \{ \hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n \} \), and a confidence score associated to those masks \( s = \{ s_1, s_2, \ldots, s_n \} \).

Given that the sequential order in the prediction does not matter, we find the optimal matching between predicted and ground truth masks. This can be found out efficiently by means of the Hungarian algorithm

\[ f_{\text{Hungarian}}(\hat{Y}, y) \]

Our objective function is:

\[ \ell(\hat{Y}, s, Y) = \min_{\hat{Y}, f} \sum_{t=1}^{n} f_{\text{IoU}}(\hat{Y}_t, Y_t) \delta_{t, t} + \lambda \sum_{t=1}^{n} f_{\text{BCE}}([t \leq n], s_t) \]

\( S \) is the set of all possible matchings between prediction and ground truth. \( f_{\text{IoU}}(\hat{Y}, y) = \frac{\text{IoU}(\hat{Y}, y)}{\text{IoU}(\hat{Y}, y) + \text{IoU}(\hat{Y}, y)} \) is a relaxed version of the intersection over union (IoU).

\( f_{\text{BCE}}(a, b) = -a \log(b) + (1 - a) \log(1 - b) \) is the binary cross entropy.

Experiments

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>[1]</th>
<th>[2]</th>
<th>[3]</th>
<th>RIS</th>
<th>RIS + CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP(^*) (6.5)</td>
<td>45.8</td>
<td>48.3</td>
<td>47.9</td>
<td>48.8</td>
<td>46.7</td>
<td>50.1</td>
</tr>
<tr>
<td>AP(^*) Ave</td>
<td>39.6</td>
<td>42.9</td>
<td>41.9</td>
<td>43.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPK</th>
<th>Not</th>
<th>MSU</th>
<th>Wog</th>
<th>PRI</th>
<th>RIS</th>
<th>RIS + CRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC ((DiC))</td>
<td>-1.9</td>
<td>-3.6</td>
<td>-2.1</td>
<td>-0.4</td>
<td>0.8</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>(2.5)</td>
<td>(2.4)</td>
<td>(1.6)</td>
<td>(3.0)</td>
<td>(1.5)</td>
<td>(1.4)</td>
<td>(1.4)</td>
</tr>
<tr>
<td>((DiC))</td>
<td>2.6</td>
<td>3.8</td>
<td>2.3</td>
<td>2.2</td>
<td>2.3</td>
<td>2.3</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td>(1.8)</td>
<td>(2.0)</td>
<td>(1.5)</td>
<td>(2.0)</td>
<td>(2.0)</td>
<td>(2.0)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>SBD ((SBD))</td>
<td>74.4</td>
<td>68.3</td>
<td>66.7</td>
<td>71.1</td>
<td>56.8</td>
<td>56.8</td>
<td>66.6</td>
</tr>
<tr>
<td></td>
<td>(4.3)</td>
<td>(6.3)</td>
<td>(7.6)</td>
<td>(6.2)</td>
<td>(8.2)</td>
<td>(8.7)</td>
<td></td>
</tr>
</tbody>
</table>

Conclusion

Learning end-to-end instance segmentation is possible by means of a recurrent neural network.

We have shown that a recurrent structure is able to track visited areas in the image as well as to handle occlusion among instances.

References

Individualness and Determinantal Point Processes for Pedestrian Detection  
Donghoon Lee¹, Geonho Cha¹, Ming-Hsuan Yang², and Songhwai Oh¹  
¹Department of ECE, ASRI, Seoul National University, Korea  
²Electrical Engineering and Computer Science, University of California at Merced  

What is the determinantal point process (DPP)?  
Given $N$ items $\{1, 2, ..., N\}$, pick subset of items that maximizes  
\[
P(\text{subset} = (i, j)) \propto \det \begin{pmatrix} q_i^2 & q_i q_j & S_{ij} \\ q_i q_j & q_j^2 & S_{ij} \\ S_{ij} & S_{ij} & 1 \end{pmatrix}  
\]  
where $q_i$ is the quality (unary term) of item $i$ and $S_{ij}$ is the similarity (pairwise term) between item $i$ and item $j$.  
Pick diverse & high quality items.  

Experimental results  
INRIA dataset (less occlusion)  
PETS dataset (frequent occlusion)  

Other approaches  
Computation time

Detection Result Using Appearance Individualness  
Spatial Feature Correlation Map  
Final Correlation Map  
Final detection

Other approaches  
Computation time

Why is it important?  
Recall of the DPM detector on the challenging PETS2009 dataset is 95%  
NMS blindly suppresses other high-scored detection candidates.  

Codes and dataset are released at: http://cpslab.snu.ac.kr/software

Please send any comments or questions to donghoon.lee@cpslab.snu.ac.kr