Peak-Piloted Deep Network for Facial Expression Recognition

Xiangyun Zhao[1] Xiaodan Liang[2] Luoqi Liu[3] Teng Li[4] Yugang Han[3] Nuno Vasconcelos[1] Shuicheng Yan[3]

1.Statistical Visual Computing Laboratory, University of California, San Diego 2. Carnegie Mellon University 3. 360 AI Institute 4. Institute of Automation, Chinese Academy of Sciences

ECCV'16 EUROPEAN CONFERENCE ON COMPUTER VISION

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.

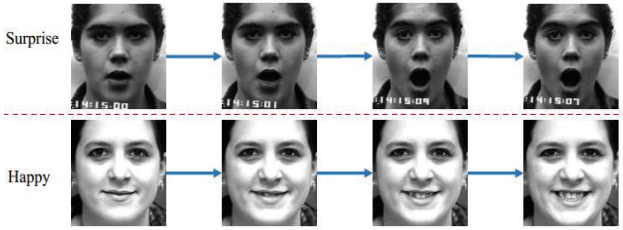


Fig1. Expression evolving process from non-peak expression to peak expression

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.

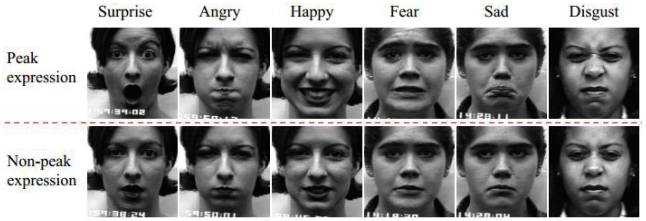
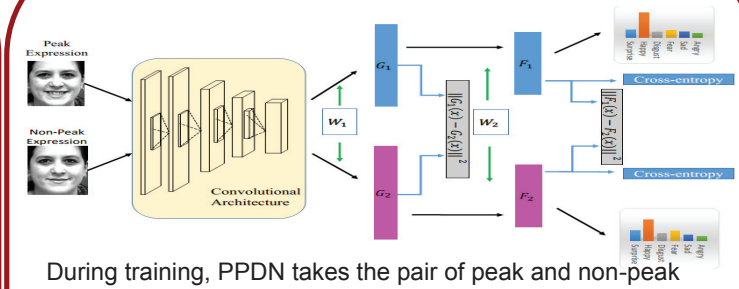


Fig2. Examples of six facial expression samples, including surprise, angry, happy, fear, sad and disgust. For each subject, the peak and non-peak expressions are shown

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$S = \{x_i^p, x_i^n, y_i^p, y_i^n, i = 1 \ldots . N\}$, where x_i^n denotes a face with non-peak expression, x_i^p a face with corresponding peak expression, and y_i^n and y_i^p 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

$$W^+ = W - \frac{\gamma}{N}\frac{\partial J_1(W; x_i^n, x_i^p)}{\partial f_j(W; x_i^n)} \times \frac{\partial f_j(W; x_i^n)}{\partial W} - \frac{1}{N}\gamma \nabla_W J_2(W; x_i^p, y_i^p) - \frac{1}{N}\gamma \nabla_W J_3(W; x_i^n, y_i^n) - 2\gamma W.$$

Where γ 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+ [1] and Oulu-CASIA [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 [3] .

Table1. Performance comparisons on six facial expressions with four state-of-the-art methods and the baseline on CK+ database.

Method	Average Accuracy
CSPL [10]	89.9%
AdaGabor [34]	93.3%
LBPSVM [11]	95.1%
BDBN [4]	96.7%
GoogLeNet(baseline)	95.0%
PPDN	**97.3%**

Table2. Performance comparisons on six facial expressions with UDCS method and the baseline on Oulu-CASIA.

Method	Average Accuracy
UDCS [35]	49.5%
GoogLeNet(baseline)	66.6%
PPDN	**72.4%**

Table3. Performance comparison on CK+ when evaluating on three different test sets, including "weak expression", "peak expression" and "combined", respectively.

Method	weak expression	peak expression	combined
PPDN(standard SGD)	81.34%	99.12%	94.18%
GoogLeNet (baseline)	78.10%	98.96%	92.19%
PPDN	**83.36%**	**99.30%**	**95.33%**

Trable4. Performance comparison on Oulu-CASIA database when evaluating on three different test sets, including "weak expression", "peak expression" and "combined", respectively.

Method	weak expression	peak expression	combined
PPDN(standard SGD)	67.05%	82.91%	73.54%
GoogLeNet (baseline)	64.64%	79.21%	71.32%
PPDN	**67.95%**	**84.59%**	**74.99%**

Table 5. Face recognition rates for various poses on Multi-PIE with some state of the arts.

Method	−45°	−30°	−15°	+15°	+30°	+45°	Average
Li et al. [38]	56.62%	77.22%	89.12%	88.81%	79.12%	58.14%	74.84%
Zhu et al. [27]	67.10%	74.60%	86.10%	83.30%	75.30%	61.80%	74.70%
CPI [28]	66.60%	78.00%	87.30%	85.50%	75.80%	62.30%	75.90%
CPF [28]	73.00%	81.70%	89.40%	89.50%	80.50%	70.30%	80.70%
GoogLeNet (baseline)	56.62%	77.22%	89.12%	88.81%	79.12%	58.14%	74.84%
PPDN	**72.06%**	**85.41%**	**92.44%**	**91.38%**	**87.07%**	**70.97%**	**83.22%**

[1] Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohnkanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, IEEE (2010) 94–101
[2] Zhao, G., Huang, X., Taini, M., Li, S.Z., Pietikainen, M.: Facial expression recognition from near-infrared videos. Image and Vision Computing 29(9) (2011) 607–619
[3] Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing 28(5) (2010) 807–813