Webly-supervised Video Recognition by Mutually Voting for Relevant Web Images and Web Video Frames

Chuang Gan Chen Sun Lixin Duan Boqing Gong
Tsinghua University Google Research Amazon University of Central Florida

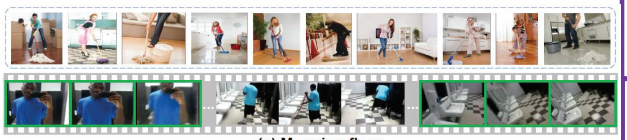
ECCV'16 EUROPEAN CONFERENCE ON COMPUTER VISION

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

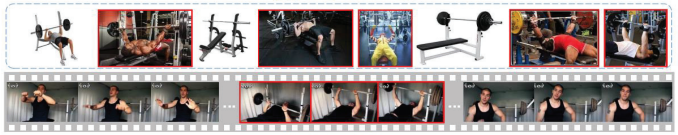


(a) Mopping floor

- ✓ 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

(a) Basketbal Dunk



(b) Bench Press

- ✓ 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.

$$\min_{\substack{\hat{\boldsymbol{\alpha}}\in[0,1]^{\mathsf{M}},\\ \hat{\boldsymbol{\beta}}\in[0,1]^{\mathsf{N}},W}} \underbrace{\left(\hat{\boldsymbol{\alpha}}^{\top},\hat{\boldsymbol{\beta}}^{\top}\right)\begin{pmatrix} K_I & -K_{IV} \\ -K_{IV}^{\top} & K_V \end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{\alpha}}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}}_{Aggressive Matching} + \underbrace{\lambda\|V - V\cdot diag(\hat{\boldsymbol{\beta}})\cdot W\|_F^2}_{Passive\ \ Selection},$$

4. Experiment

Action recognition on UCF101

Method	# Number of training data	Acc (%)
All crawled data	426K	64.7
Validation	368K	66.5
One-class SVM (10%)	384K	65.9
One-class SVM (15%)	363K	65.9
Unsupervised One-class SVM (10%)	384K	66.9
Unsupervised One-class SVM (15%)	363K	66.4
Landmarks (10%)	384K	68.3
Landmarks (15%)	363K	67.7
Ours (10%)	384K	69.3
Ours (15%)	363K	68.9

Video event detection on TRECVID MED 2013

Method	mAP (%)
Composite concept	6.4
EventNet	8.9
Selecting	11.8
Ours	16.1

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.