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.

\[
\min_{\alpha \in [0, 1]^M, \beta \in [0, 1]^m, W} \left( \alpha^T \beta^T \left( \begin{pmatrix} K_T & -K_{TV} \\ -K_{TV}^T & K_V \end{pmatrix} \right) \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right) + \lambda \| V - V \cdot \text{diag}(\beta) \cdot W \|_F^2 \]  

4. Experiment

- Action recognition on UCF101
- Video event detection on TRECVID MED 2013

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.