Discovering Objects in Videos

Karim Sayed Ahmed
Dartmouth College
Department of Computer Science
karim@cs.dartmouth.edu

Machine Learning - Project Proposal.

Problem

The growing amount of videos requires automatic and intelligent approaches that aim at analyzing and understanding the visual data which in consequence help in improving the effectiveness and efficiency of videos. One of these approaches is "discovering objects in videos", which is useful for many applications and can be utilized for many higher-level vision tasks such as activity recognition, video summarization, and visual enhancement.

The goal of this project is to identify and localize objects in videos. Given a video V, consisting of a set of N frames $\{f_1, . . , f_N\}$ where each frame contains one or more unknown object (unknown category/shape + unknown location), I want to identify and localize **primary** object(s) in each i**-th** frame f_i in V that exhibit coherence in both appearance and motion. **Figure 1** shows a demo for the final output.

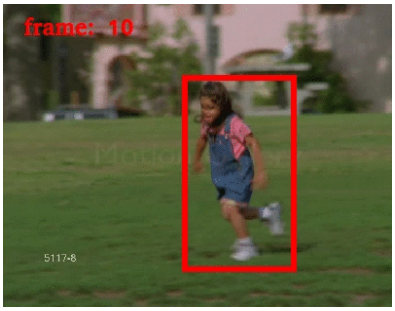

Figure 1: Demo for final output of "discovering objects in videos". This short video consists of 21 frames and shows a moving girl (the primary object). The red box marks the girl in each frame. The original video frames taken from website [5], and annotated manually by red color.

Method

Extending object localization techniques in still images to be used in videos. The proposed method operates in offline mode and consists of **two main steps** (2 learning problems):

Step 1) Object localization per frame using deep learning:

For every i-th video frame f_i (still image), generate bounding box b_i**(**x, y, w, h**)** that is very likely to contain primary object, where x: x-coordinate of bounding box b_i, y: y-coordinate of bounding box b_i, w: width of bounding box b_i, h: height of bounding box b_i.

For the generating the bounding box for each video frame that is most likely contains a primary object, I will use the approach proposed in [1]. This approach is a new self-taught object localization algorithm for still images that leverages on **deep convolutional neural networks** trained for whole image recognition to localize objects in images without additional human supervision, i.e., without using any ground-truth bounding boxes for training [1]. This approach can be used to generate as many bounding boxes as needed (containing multiple candidate objects in same image); however I will select the highest-scoring bounding box only for each video frame; this highest-scoring bounding box is most likely to contain the primary object. **Figure 2** shows the an example for the output of this step.

Figure 2: Example of results in Step 1, red bounding box in still images shows the primary objects identified and localized using approach in [1]. Figure from paper [1]

Step 2) Classification of primary objects from all video frames to find correlation between them:

After identifing and localizing the primary object in each frame (output of step 1), I will use a suitable classifier to find the correlation between the generated bounding boxes (objects) in different video frames. This learning problem could be formulated as detection of similarity between objects; discarding noise objects. Within the same video, similar objects in different frames are most likely to be the same object moving through time.

The are two main issues with learning the similarity between frames of the same video which are:

- Few labeled samples in training data especially for positive samples.
- The testing data are part of the training data.

To overcome these issues, I propose to use **Transductive Support Vector Machines** described in [3], in a similar manner as used in [2]. The benefit of using Transductive Support Vector Machines is that it allows using a mixture of labeled and unlabeled samples as training data, and also the unlabeled samples used in training stage are used in testing stage.

In conclusion, the proposed plan for step 2 is to use Transductive Support Vector Machines [3], applied on the primary objects generated by step 1, where each primary object is represented by the rich object features described in [4].

Dataset and Evaluation

I plan to use **"SegTrack"** dataset [**5, 6**][website], which provides quantitive evaluation. The following are characteristics of this dataset:

- Dataset of 6 video sequences
- Provides pixel-level segmentation ground-truth for each video
- Performance metrics: average per-frame pixel error compared to the ground-truth
- Widely used by many other approaches in the litrature for example [**5, 6, 7, 8**], so the proposed approach can be compared to them as well.

Timeline

By milestone due date (October 28, 2014),

- I would finish generating bounding boxes for videos in the dataset (Step 1).
- I would finish refining existing code or adding new components to existing code.
- I would have started implementing the classifier (Step 2).

References

[1] Alessandro Bergamo, Loris Bazzani, Dragomir Anguelov, and Lorenzo Torresani. "Self-taught Object Localization with Deep Networks". arXiv:1409.3964, 2014.

[2] Alessandro Bergamo, Lorenzo Torresani. "Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach". In NIPS 2010.

[3] Thorsten Joachims. "Transductive inference for text classification using support vector machines". In ICML 1999.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". arXiv:1311.2524, 2013.

[5] David Tsai, Matthew Flagg and James M.Rehg,"Motion Coherent Tracking with Multi-label MRF Optimization", BMVC 2010. [Website]

[6] David Tsai, Matthew Flagg, Atsushi Nakazawa, and James M. Rehg. "Motion coherent tracking using multi-label mrf optimization". International journal of computer vision (IJCV) 100, Vol no. 2 (2012): 190-202.

[7] Yong Jae Lee, Jaechul Kim, and Kristen Grauman. "Key-segments for video object segmentation". In ICCV 2011.

[8] Dong Zhang, Omar Javed, and Mubarak Shah. "Video object segmentation through spatially accurate and temporally dense extraction of primary object regions." In CVPR 2013.