Multiple Frames Matching for Object Discovery in Video

Otilia Stretcu
otiliastr@gmail.com
Computer Laboratory
University of Cambridge, UK

Marius Leordeanu
marius.leordeanu@imar.ro
Institute of Mathematics of the Romanian Academy
Teamnet International, Romania

Automatic discovery of foreground objects in video sequences is important in computer vision, with applications to object tracking, video segmentation and weakly supervised learning. This task is related to co-segmentation [4, 5] and weakly supervised localization [2, 6]. We propose an efficient method for the simultaneous discovery of foreground objects in video and their segmentation masks across multiple frames. We offer a graph matching formulation for bounding box selection and refinement using second and higher order terms. It is based on an Integer Quadratic Programming formulation and related to graph matching and MAP inference [3]. We take into consideration local frame-based information as well as spatiotemporal and appearance consistency over multiple frames. Our approach consists of three stages. First, we find an initial pool of candidate boxes using a novel and fast foreground estimation method in video (VideoPCA) based on Principal Component Analysis of the video content. The output of VideoPCA combined with Edge Boxes [8] is then used to produce high quality bounding box proposals. Second, we efficiently match bounding boxes across multiple frames, using the IPFP algorithm [3] with pairwise geometric and appearance terms. Third, we optimize the higher order terms using the Mean-Shift algorithm [1] to refine the box locations and establish appearance regularity over multiple frames. We make the following contributions:

1. A novel formulation with efficient discrete and continuous optimization for joint selection and refinement of object bounding boxes in video. Our approach encourages appearance, geometric and spatiotemporal consistency over multiple frames with a formulation that considers relations between neighboring as well as farther away frames (Figure 1).

2. A fast method for estimating foreground regions based on Principal Component Analysis of the video content that is able to handle cases of moving or changing backgrounds.

Given a video V as a sequence of temporally ordered frames $V = \{I_1, I_2, ..., I_n\}$, the goal is to find the main foreground object present in the sequence. For each frame I_i we have a pool of n_i candidate bounding boxes B_{ia}'s, obtained automatically. For each box a we store its xy location in image i in θ_{ia}. Let x_{ia} be an indicator variable corresponding to bounding box B_{ia} (- the a-th bounding box of frame i), such that x_{ia} is 1 if the bounding box B_{ia} is selected, and 0 otherwise. The indicator variables are arranged in vector form $\mathbf{x}$ such that its ia-th element corresponds to x_{ia}. We impose the constraint that a single box can be selected per frame: $\sum_a x_{ia} = 1$. Thus, vector $\mathbf{x}$ represents a discrete solution that indicates which box is selected. Similarly, we keep the continuous location parameters in a global vector θ, with θ_{ia} being the parameters of box B_{ia}. The problem becomes one of joint box matching and location refinement over multiple frames, in which we optimize over both $\mathbf{x}$ and θ:

$$\begin{aligned} (\mathbf{x}^*, \theta^*) &= \underset{x,\theta}{\operatorname{argmax}} (\mathbf{x}^T \mathbf{M} \mathbf{x} + \sum_{i=1}^{n} H_i(\mathbf{x}, \theta)) \\ \text{s.t.} \quad & \sum_a x_{ia} = 1 \ \ \forall i, \ \ \mathbf{x} \in \{0,1\}^n. \end{aligned}$$

We adopt a two stage approach. First stage performs discrete optimization in which the quadratic function is optimized by finding the correct frame-to-box matches $\mathbf{x}$, given a fixed θ. The second stage, when $\mathbf{x}^*$ is fixed, the location θ is refined by the non-parametric Mean-Shift to locally optimize the foreground pixels density.

The **pairwise terms** have the following form:

$$M_{ia;jb} = \exp(\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb}). \quad (1)$$

$\mathbf{g}_{ia;jb} = [(f_{ia} + f_{jb}), (v_{ia} + v_{jb}), (c_{ia} + c_{jb}), m_{ia;jb}, o_{ia;jb}, d_{ia;jb}, s_{ia;jb}, r_{ia;jb}]$, such that: 1) f_{ia} (and f_{jb} respectively) measure the absolute difference between the average foreground soft-segmentation values in box a (and b respectively) vs. average foreground values in the surrounding background from its frame i (and j respectively). 2) Similarly, v_{ia} and v_{jb} measure the absolute difference in the relative mean speed between the box and the surrounding background, computed using the DeepFlow method [7]. 3) c_{ia} and c_{jb} measure the distance between the box center and the image center. 4) $m_{ia;jb}$ reflects the quality of the match between the standard HOG descriptor of box a and that of box b. 5) $o_{ia;jb}$ measures the overlap-over-union between the boxes. 6) $d_{ia;jb}$ is the distance between the boxes' centers. 7) $s_{ia;jb}$ is the ratio of the difference between the boxes' areas to the maximum of the two areas. 8) $r_{ia;jb}$ estimates the change in shape, as difference between the boxes' aspect ratios. We learn parameters $\mathbf{w}$ such that $\exp(\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb})$ approximates a target $t = 1$ if the matched pair is correct and equal to a small positive value ($t = 0.1$) otherwise. We want $\exp(\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb}) \approx t$. We take the log on both sides $\mathbf{w}^\mathbf{T} \mathbf{g}_{ia;jb} \approx \log t$ and obtain a linear system of equations over a set of training samples, which we approximately solve by ridge regression.

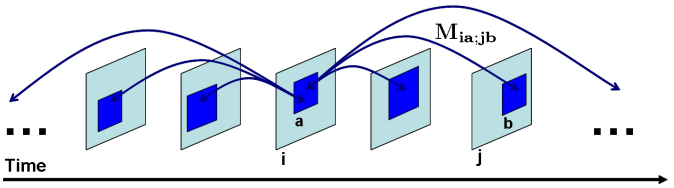

Figure 1: The structure of our box-matching formulation allows links between neighboring frames as well as farther away ones.

For the **higher order terms** we estimate foreground and background color probability distributions from the $2k+1$ bounding boxes and their frames connected to the current one (including itself) and estimate the soft foreground segmentation by classifying individual pixels based on their color. The higher order term $H_i(\mathbf{x}, \theta) = \lambda c_k(i)$ measures the difference between average foreground segmentation values inside the box and outside of it.

We test our method on the large scale YouTube-Objects dataset and obtain state-of-the-art results on several object classes, at a low computational cost. We conclude that our approach, by proposing efficient bounding box generation with VideoPCA, fast matching over multiple frames, and high quality soft-segmentation estimation, covers and extends current approaches in object discovery in video.

[1] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence*, 24(5), 2002.

[2] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3), 2012.

[3] M. Leordeanu, M. Hebert, and Rahul Sukthankar. An integer projected fixed point method for graph matching and map inference. In *NIPS*, 2009.

[4] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[5] J.C. Rubio, J. Serrat, and A. López. Video co-segmentation. In *ACCV*, 2012.

[6] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013.

[7] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *ICCV*, 2013.

[8] C. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.