Semantic Object Segmentation via Detection in Weakly Labeled Video

Yu Zhang[1], Xiaowu Chen[1*], Jia Li[1,2], Chen Wang[1], Changqun Xia[1]
[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University.
[2]International Research Institute for Multidisciplinary Science, Beihang University. (*corresponding author: chen@buaa.edu.cn)

Semantic video object segmentation is an essential step in computer vision and multimedia analysis. However, in many cases, semantic objects are only weakly tagged at video-level, making them difficult to be located and segmented. Recently, several approaches have been proposed to address video segmentation in such a setting through weakly supervised learning [3, 4, 5]. Although having succeeded in certain scenarios, these approaches may suffer the ambiguity of training instances and thus generate unexpected results. Moreover, multiple videos are required by these approaches during training, which may prevent their usages in single video segmentation.

Differentiate from previous studies, a segmentation-by-detection framework is presented in this paper for semantic object segmentation in weakly labeled video. The proposed framework makes use of the power of image-based object detectors and thus avoids the ambiguous training procedure. In this framework, image-based object detectors [1, 2] are first employed on various frames to generate a set of detection/segmentation proposals, which however may be noisy and lack spatiotemporal consistency. Therefore, a joint assignment problem is proposed and efficiently solved to initialize object tracks from noisy proposals. The initial tracks are then refined through spatiotemporally consistent shape likelihoods inferred from statistical information of segment tracks. We briefly introduce these steps as follows.

Given a video with T frames, a set of object detection proposals $\mathbb{D}_t$ and segmentation proposals $\mathbb{S}_t$ are extracted for the tth frame. In the track initialization, the objective is to find K object tracks that best cover the semantic objects from the noisy proposals, which can be achieved by jointly assigning detections and segmentations to tracks. Define the binary variables

$$\mathbb{A} = \{a^k_{\mathcal{D}} | \forall k, t, \mathcal{D} \in \mathbb{D}_t\}, \;\; \mathbb{B} = \{b^k_{\mathcal{S}} | \forall k, t, \mathcal{S} \in \mathbb{S}_t\}, \tag{1}$$

where $a^k_{\mathcal{D}}, b^k_{\mathcal{S}} \in \{0,1\}$ represent the assignment of detection $\mathcal{D}$ and segmentation $\mathcal{S}$ to the kth track, respectively. We make the following assumptions: 1) a track selects at most one segmentation or detection on a frame, 2) each selected segmentation should be coupled with a detection, 3) tracks are non-overlapping, 4) consecutive and 5) not empty. Under these assumptions, we optimize $\mathbb{A}$ and $\mathbb{B}$ by solving

$$\begin{aligned}
&\min_{\mathbb{A},\mathbb{B}} \mathcal{L}(\mathbb{A},\mathbb{B}) + \lambda_1 \Omega_1(\mathbb{A},\mathbb{B}) + \lambda_2 \Omega_2(\mathbb{B}), \\
&\text{s.t. } a^k_{\mathcal{D}}, b^k_{\mathcal{S}} \in \{0,1\}, \; \forall k, t, \mathcal{D} \in \mathbb{D}_t, \mathcal{S} \in \mathbb{S}_t, \\
&\quad \sum_{\mathcal{D} \in \mathbb{D}_t} a^k_{\mathcal{D}} = \sum_{\mathcal{S} \in \mathbb{S}_t} b^k_{\mathcal{S}} \le 1, \; \forall k, t, \\
&\quad \sum_k a^k_{\mathcal{D}} \le 1, \sum_k b^k_{\mathcal{S}} \le 1, \; \forall t, \mathcal{D} \in \mathbb{D}_t, \mathcal{S} \in \mathbb{S}_t, \\
&\quad \left(\sum_{\mathcal{D} \in \mathbb{D}_{t-p}} a^k_{\mathcal{D}}\right) \left(1 - \sum_{\mathcal{D} \in \mathbb{D}_t} a^k_{\mathcal{D}}\right) \left(\sum_{\mathcal{D} \in \mathbb{D}_{t+q}} a^k_{\mathcal{D}}\right) = 0, \forall k, t, p, q, \\
&\quad \sum_t \sum_{\mathcal{D} \in \mathbb{D}_t} a^k_{\mathcal{D}} = \sum_t \sum_{\mathcal{S} \in \mathbb{S}_t} b^k_{\mathcal{S}} \ge 1, \forall k, t,
\end{aligned} \tag{2}$$

where $\mathcal{L}$ is the loss function, Ω_1 and Ω_2 penalize the tracking inconsistency and track interactions, respectively. We omit the definition of $\mathcal{L}$, Ω_1 and Ω_2 in this abstract and refer the reader to the full paper for details. The problem (2) is a constrained combinatorial optimization problem on massive binary variables, which is hard to optimize directly. However, as shown in the paper, it can be reformulated into a min-cost flow problem and heuristically solved through a two-step algorithm. In the first step, several independent object tracks are efficiently initialized, and then refined in the second step that takes the mutual overlaps of tracks into consideration.

The initial tracks can roughly locate the semantic objects in the video, but may have inconsistent appearance in different frames. For each initial track, this paper further proposes to improve its visual coherence by estimating spatiotemporally consistent shape likelihoods. To this end, a series of N segment tracks that overlap with the initial track are extracted by greedily linking the pre-segmentations across frames. A set of scores $\{\alpha^0_i\}^N_{i=1}$ are then computed as an initial estimation of the confidence of the N tracks. As discussed in the paper, an optimization stage is conducted on the track scores instead of pixel-level inference, through which the higher-order object properties can be leveraged. Formally, the adjusted scores $\boldsymbol{\alpha} = \{\alpha_i\}^N_{i=1}$ are obtained through solving

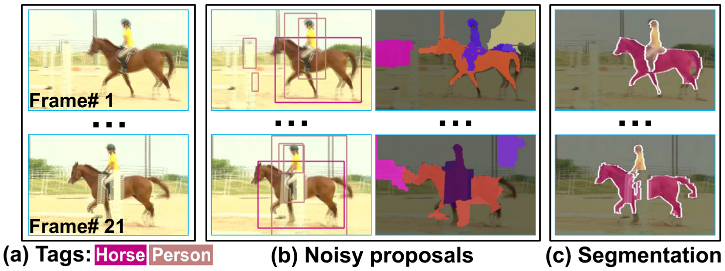

Figure 1: The motivation of our approach. (a) The input video is weakly labeled with semantic tags, making it difficult to locate and segment the desired objects (e.g., the horse is occluded by fence in frame #21); (b) Per-frame detection and segmentation proposals provide location information but are often very noisy; (c) The proposed segmentation-by-detection framework can generate consistent object segmentation results from noisy detection and segmentation proposals.

$$\min_{0 \preceq \boldsymbol{\alpha} \preceq 1} \sum_{i=1}^{N} \left(\alpha_i - \alpha^0_i\right)^2 + \theta_1 \mathcal{C}_1(\boldsymbol{\alpha}) + \theta_2 \mathcal{C}_2(\boldsymbol{\alpha}), \tag{3}$$

where the first term penalizes the deviation of the adjusted scores from initial estimation, the second and the third terms account for appearance and temporal consistency, respectively. As detailed in the paper, the appearance consistency term allows visually similar tracks to have similar scores, while the temporal consistency term enforces shape likelihoods of temporally adjacent pixels to change slowly. The initial tracks are finally refined using the inferred shape likelihoods via a graph-cut based optimization procedure [6].

With the proposed framework, improvements over several state-of-the-art weakly supervised and unsupervised approaches are observed on both Youtube-Objects and SegTrack v2 datasets. This paper makes several contributions, including: 1) a novel segmentation-by-detection framework for semantic object segmentation in weakly labeled video, 2) an algorithm to robustly initialize object tracks from noisy proposals by solving a joint assignment problem, and 3) an inference step of spatiotemporally consistent shape likelihoods from the statistical information of segment tracks.

[1] J. Carreira and C. Sminchisescu. CPMC: Automatic object segmentation using constrained parametric min-cuts. *PAMI*, 34(7):1312–1328, 2012.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[3] G. Hartmann, M. Grundmann, J. Hoffman, D. Tsai, V. Kwatra nad O. Madani, S. Vijayanarasimhan, I. Essa, J. Rehg, and R. Sukthankar. Weakly supervised learning of object segmentations from web-scale video. In *ECCV Workshop*, 2012.

[4] X. Liu, D. Tao, M. Song, Y. Ruan, C. Chen, and J. Bu. Nearest neighbor-based label transfer for weakly supervised multi class video segmentation. In *CVPR*, 2014.

[5] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013.

[6] W. Xia, C. Domokos, J. Dong, L.-F. Cheong, and S. Yan. Semantic segmentation without annotating segments. In *ICCV*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.