JOTS: Joint Online Tracking and Segmentation

Longyin Wen[1], Dawei Du[2], Zhen Lei[1], Stan Z. Li[1], Ming-Hsuan Yang[3]
[1]Natinal Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences.
[2]School of Computer and Control Engineering, University of Chinese Academy of Sciences.
[3]School of Engineering, University of California at Merced.

Despite much demonstrated success, existing methods are less effective for applications that entail online processing. Examples abound, including video surveillance, action recognition and human-computer interaction, to name a few. Recently, some online video segmentation methods are proposed, e.g., [4] and [6]. The global object appearance modeling without strong local constraints in [4] and the target-independent proposals generation step in [6] may cause inaccurate segmentation results, especially in the scene with complex background or large motions.

We present a novel Joint Online Tracking and Segmentation (JOTS) algorithm which integrates the multi-part tracking and segmentation into a unified energy optimization framework to handle the video segmentation task. The multi-part segmentation is posed as a pixel-level label assignment task with regularization according to the estimated part models, and tracking is formulated as estimating the part models based on the pixel labels, which in turn is used to refine the model. The multi-part tracking and segmentation are carried out iteratively to minimize the proposed objective function by a RANSAC-style approach.

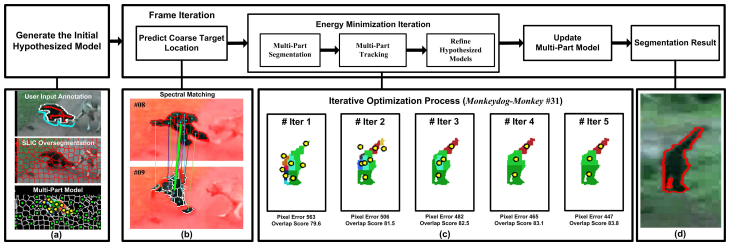

Figure 1: Main steps (upper part) and corresponding example (lower part) of the proposed JOTS algorithm. (a) Initial multi-part model construction. (b) Spectral matching for generating approximate target location. (c) An example showing the iterative optimization process. (d) Final segmentation result where the target boundary is delineated by red pixels.

Given simple user annotation followed by the interactive segmentation method in the first frame, we first segment a target object from the background, and then use the Simple Linear Iterative Clustering (SLIC) algorithm to generate the initial hypothesized models, as depicted in Figure 1(a). The video segmentation task in this work is formulated by multi-part tracking and segmentation in a unified framework. That is, we optimize the pixel labels $f \in \{l_0\} \cup \{l_1, \ldots, l_k\}$ and the target multi-part model $\mathcal{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$ simultaneously. For each pixel p in the image, we assign it with a label f_p indicating which part it belongs to (i.e., multi-part segmentation) rather than merely identify it as the foreground or background in previous methods, and optimize the target multi-part model $\mathcal{M}$ in the current frame (i.e., multi-part tracking) simultaneously. The video segmentation problem is formulated as $\{\mathcal{M}^*, f^*\} = \text{argmin}_{\mathcal{M},f} E(\mathcal{M}, f|\overline{\mathcal{M}})$, where $\overline{\mathcal{M}}$ is the multi-part model in the previous frame, and $\mathcal{M}^*$ and f^* are the multi-part model and the pixel labeling result in the current frame. To solve this problem, we first obtain the predicted target location using the dynamic structure graph matching method [3], as shown in Figure 1(b). The coarsely estimated target location is computed based on the votes of the matched parts and determines the segmentation region.

In the segmentation region, we compute the solution $\{\mathcal{M}^*, f^*\}$ by minimizing $E(\mathcal{M}, f|\overline{\mathcal{M}})$. For the clarity of presentation, we omit $\overline{\mathcal{M}}$ in the following equations. Then a unified objective function integrates both multi-part tracking and segmentation information, defined as

$$\{\mathcal{M}^*, f^*\} = \underset{\mathcal{M},f}{\text{argmin}} \left\{ D(f, \mathcal{M}) + \sum_{(p,q)\in\mathcal{N}} V_{p,q}(f_p, f_q) + H(f, \mathcal{M}) \right\}, \quad (1)$$

where $D(f, \mathcal{M})$ is the data term based on the current labeling f and multi-part model $\mathcal{M}$, $V_{p,q}(f_p, f_q)$ is the smooth term describing the interactions between neighboring pixels, and $H(f, \mathcal{M})$ is the regularization term of $D(f, \mathcal{M})$ to avoid overfitting [1, 5] by enforcing constraints of the models in pixel labeling.

A RANSAC-style method is proposed to obtain the solution in two steps (See Figure 1(c)): 1. the pixel labels are assigned with the current estimated multi-part model by the α-expansion algorithm [2]; 2. the target parts are tracked according to the pixel appearance likelihood and motion consistency with the current labeling.

Multi-part segmentation. In the multi-part segmentation stage, pixel labels f are computed by the α-expansion algorithm with the regularization term and a small set of reliable models in $\mathcal{M}$ are selected. The pixel labeling problem is formulated as an energy minimization of the pairwise Markov random field,

$$f^* = \underset{f}{\text{argmin}} D(f, \mathcal{M}) + \sum_{(p,q)\in\mathcal{N}} V_{p,q}(f_p, f_q) + H(f, \mathcal{M}), \quad (2)$$

where $D(f, \mathcal{M})$ is the data term based on the labeling f and multi-part model $\mathcal{M}$, $V_{p,q}(f_p, f_q)$ is the smooth term describing interactions between neighboring pixels, and $\mathcal{N}$ is the 4-neighborhood relations between pixels. The optimization problem can be solved by the α-expansion algorithm [2] with graph cut effectively.

Multi-part tracking. In the multi-part tracking stage, some selected models are improved by re-estimating multi-part model with the energy function (3). Next, we add some hypothesized part models based on the current labeling to expand the multi-part model $\mathcal{M}$. Given the current pixel labels f, the multi-part tracking problem is formulated as

$$\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmin}} D(f, \mathcal{M}) + H(f, \mathcal{M}), \quad (3)$$

Similar to [1], we disregard the regularization term at first and focus on minimizing the first term of (3) using the Maximum Likelihood Estimation (MLE) method to obtain the optimal models $\mathcal{M}^*$.

These two steps are iterated until reaching the minimal energy of the objective function such that the multi-part tracking facilitates the multi-part segmentation, and vice versa. After the iterative optimization, we update the multi-part model based on the optimal labeling and output the final segmentation result (See Figure 1(d)).

[1] Anton Andriyenko, Konrad Schindler, and Stefan Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR*, pages 1926–1933, 2012.

[2] Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[3] Zhaowei Cai, Longyin Wen, Zhen Lei, Nuno Vasconcelos, and Stan Z. Li. Robust deformable and occluded object tracking with dynamic graph. *TIP*, 23(12):5497–5509, 2014.

[4] J. Chang and J. W. Fisher III. Topology-constrained layered tracking with latent flow. In *ICCV*, Dec 2013.

[5] Andrew Delong, Anton Osokin, Hossam N. Isack, and Yuri Boykov. Fast approximate energy minimization with label costs. *IJCV*, 96(1): 1–27, 2012.

[6] Fuxin Li, Taeyoung Kim, Ahmad Humayun, David Tsai, and James M. Rehg. Video segmentation by tracking many figure-ground segments. In *ICCV*, 2013.

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