15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Unsupervised Video Object Segmentation by Supertrajectory Labeling

Masahiro Masuda
Waseda University
masahi129@gmail.com

Yoshihiko Mochizuki
Waseda University
motchy@aoni.waseda.jp

Hiroshi Ishikawa
Waseda University
hfs@waseda.jp

Abstract

We propose a novel approach to unsupervised video segmentation based on the trajectories of Temporal Superpixels (TSPs). We cast the segmentation problem as a trajectory-labeling problem and define a Markov random field on a graph in which each node represents a trajectory of TSPs, which we minimize using a new two-stage optimization method we developed. The adaption of the trajectories as basic building blocks brings several advantages over conventional superpixel-based methods, such as more expressive potential functions, temporal coherence of the resulting segmentation, and drastically reduced number of the MRF nodes. The most important effect is, however, that it allows more robust segmentation of the foreground that is static in some frames. The method is evaluated on a subset of the standard SegTrack benchmark and yields competitive results against the state-of-the-art methods.

1 Introduction

Successful processing and analysis of videos enables applications such as video retrieval [1], action recognition [2], and video summarization [3]. Segmentation of video is important for such higher-level tasks. Unsupervised segmentation, which requires no user intervention, is more desirable than the supervised approach in which the user manually annotates the first frame [4, 5, 6, 7]. However, it is more difficult since the segmentation must be done based on motion cues alone. Several approaches to unsupervised video segmentation have been proposed [8, 9, 10, 11]. A common preprocessing step is superpixel segmentation, usually independently applied to each frame. Superpixels are used for efficiency, to aggregate local statistics, and to compute feature descriptors. However, superpixels can be inconsistent between consecutive frames, making it hard to establish a correspondence between the frames. Since consecutive frames in a video are similar to some extent, superpixel segmentation should respect temporal coherency as well. This is exactly the goal of Temporal Superpixel (TSP) [12].

Here, we propose a new approach to unsupervised video segmentation based on the trajectories of TSPs. We cast the segmentation problem as a trajectory labeling problem, i.e., each trajectory is determined to be foreground or background. We define a Markov random field on a graph in which each node represents a trajectory of TSPs, with a novel energy which we minimize using a new two-stage optimization method we developed. The method is evaluated on a subset of the standard SegTrack benchmark and yields competitive results against the state-of-the-art methods.

2 Trajectory-based Video Segmentation

Our approach to the unsupervised video object segmentation is based on the Temporal Superpixel (TSP) [12], which produces superpixels that are temporally consistent. The TSP is a generative probabilistic model of superpixel appearance and motion that can be interpreted as a probabilistic and temporal extension of the state-of-the-art SLIC superpixel method [13].

2.1 TSP Trajectory Markov Random Field

A pair of temporal superpixels in consecutive frames are assigned the same ID if one of the superpixels is successfully tracked from the other based on color similarity and optical flow. Therefore, a set of TSPs with the same ID can be intuitively thought to form a spatio-temporal tube. It will be referred to as a *TSP trajectory* or simply *trajectory* hereafter. A trajectory can consist of a superpixel in a single frame or it can span the entire length of the video. See Fig. 1 for illustration.

Taking advantage of the remarkable stability of TSPs across frames, we assume that they do not cross the object-background boundary and formulate the video segmentation as a trajectory labeling problem. Thus, each TSP trajectory is assumed to be completely contained in the foreground or the background. In our formulation, each trajectory is a single node in an MRF.

The adaption of the TSP trajectories as basic building blocks brings several advantages over conventional superpixel-based methods [9, 8]. First, compared to a single superpixel, a sequence of superpixels enables us to design more expressive potential functions. Second, a single trajectory is temporally coherent by design, which promotes temporal coherence of the resulting segmentation. Third, using trajectory helps segment a part of the foreground that is static in some frames: in previous approaches, it was often difficult to segment a foreground that stays static for a few frames, because motion cues such as optical flow are temporally local; with the TSP algorithm, the static region can be easily tracked across many frames till it starts moving. Using the proposed unary potential described next, lack of motion in some frames can be compensated by non-static regions in other frames on the same trajectory. As an additional benefit, the number of nodes in the MRF is drastically reduced.

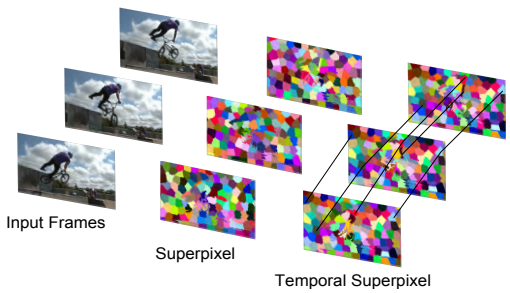


Figure 1: The Temporal Superpixel (TSP). Unlike the SLIC superpixel algorithm [13], which is applied to each frame individually (center), the TSP algorithm [12] tracks superpixels corresponding to the same regions across frames, forming trajectories of superpixels (right).

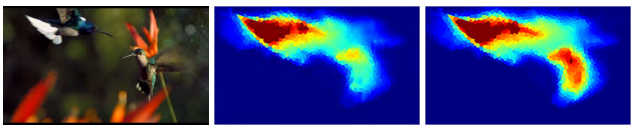


Figure 2: The location prior computed with the method of [9] (center) and with the modified method (right). Foreground likelihood is color-coded. The brighter regions are more likely to be foreground.

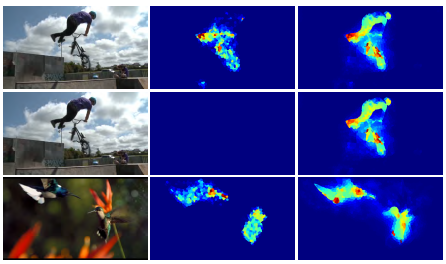


Figure 3: Diffusion results. Input frames (left), the original inside probabilities (center), and the diffused inside probabilities (right) are shown.

Let $\mathcal{V}$ the set of trajectories, and $\mathcal{E}$ the set of pairs of neighboring trajectories. Two trajectories are neighbors if they are adjacent in at least one frame. The energy is of the following form:

$$E(\mathbf{l}) = \sum_{u \in \mathcal{V}} U_u(l_u) + \sum_{(u,v) \in \mathcal{E}} V_{u,v}(l_u, l_v), \qquad (1)$$

where $\mathbf{l} = (l_u)_{u \in \mathcal{V}}$ is the *trajectory labeling*, which assigns a label $l_u \in \{0, 1\}$ to each trajectory u in $\mathcal{V}$. We let 0 represent the background and 1 the foreground.

2.2 Unary Potential

The unary potential estimates the likelihood of trajectories being foreground or background based on various cues. It consists of an appearance model $A_u(l_u)$ and a location model $L_u(l_u)$:

$$U_u(l_u) = A_u(l_u) + L_u(l_u). \qquad (2)$$

We build on the *inside-outside map* developed in [9] to compute both the appearance and the location model.

Inside-outside Map. The inside-outside map gives each pixel 1 if it is inside of the boundary and 0 otherwise. In [9], it is computed in the following steps: (1) Approximate the boundaries of the foreground by contour fragments, which are pixels with optical flow edge strength above a fixed threshold. That is, pixels in contour fragments are those with large change in optical flow, typically on the boundary between a moving object and the background. (2) Determine if each pixel is inside or outside of the boundaries by shooting rays to eight directions from the pixel. If more than four rays intersect contour fragments an odd number of times, that pixel is deemed inside of the boundary.

In [9], the contour fragments are generated by thresholding optical flow edge strength at a fixed value. We observed that for some videos this threshold is too high and weaker optical flow edge is lost after thresholding. To make the construction of the inside-outside map more robust, we select N evenly spaced values in range $[\tau_{\min}, \tau_{\max}]$ as thresholds, compute a map for each threshold, and take the average of the maps for each pixel. Thus, the *robust* inside-outside map is not binary valued, but takes values in the range $[0, 1]$. Although it is not a radical change, the resulting location model does improve, as shown in Fig. 2.

Inside Probability. The inside-outside map is defined on *pixels*. Now, for each *superpixel*, we define the *inside probability* as the ratio of pixels in that superpixel that are inside. In the robust case, it is the average of the robust inside-outside map. It can be interpreted as the probability of a superpixel being inside of the foreground boundary.

Inside Probability Diffusion. Since the inside probability is computed in each frame using optical flow, it fails when a part of the foreground is static in some frames. Therefore, we propose to diffuse the inside probability within each trajectory. The aim is to compensate the lack of motion in some frames where a part of foreground is static and thus the corresponding inside probability is small. Consider all the superpixels in the video and let p_i and d_i be the original and diffused inside probabilities of superpixel i. The diffusion is done by minimizing the convex quadratic energy

$$\sum_i k_i (d_i - p_i)^2 + \lambda \sum_i \sum_{j \in \mathcal{N}(i)} \frac{1}{2} w_{i,j} (d_i - d_j)^2,$$

where k_i are per-node weights and $w_{i,j}$ are defined as $w_{i,j} = \exp\left(-\|\mathbf{c}_i - \mathbf{c}_j\|^2 / 2\sigma^2\right)$, where $\mathbf{c}_i$ is the mean RGB color of superpixel i and σ is fixed to 30. To diffuse the inside probability within trajectories, the neighborhood $\mathcal{N}(i)$ of superpixel i is defined to be the set of superpixels that are spatially adjacent to i in the same frame or that are on the same trajectory as i. The parameter λ controls the amount of diffusion, and is set to 10 in all experiments. The minimization can be achieved by solving a sparse linear system. Figure 3 shows the diffused inside probability. Clearly, the diffusion helps uncover foreground regions with small inside probability.

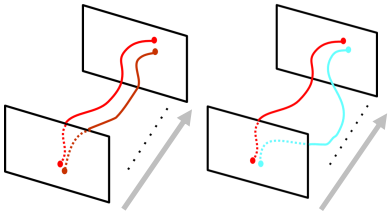
Appearance model. The appearance model is given by a random forest classifier trained with the RGB color as features. We choose training samples from the superpixels according to the diffused inside probability. For each frame, we build a histogram with 20 bins of diffused inside probability for the superpixels in the frame. Then we take the superpixels in the first nonempty bin (with the smallest diffused inside probabilities) as background samples. Skipping the next three bins, we take the superpixels in the fifth and all following bins as the foreground samples.

Let S_u be the set of superpixels in trajectory u and P^i_{app} the probability predicted by the random forest that superpixel i is foreground. Then $A_u(l_u)$ for trajectory u is defined as

$$A_u(0) = \frac{1}{|S_u|} \sum_{i \in S_u} -\log(P^i_{\text{app}}), \qquad (3)$$

$$A_u(1) = \frac{1}{|S_u|} \sum_{i \in S_u} -\log(1 - P^i_{\text{app}}). \qquad (4)$$

Location model. For the location model, the diffused inside probability is propagated forward and backward in time using the method of [9]. Let P^i_{loc} be the propagated inside probability. Then $L_u(l_u)$ for trajectory u is defined



(a) high similarity (b) low similarity

Figure 4: Two neighboring trajectories with similar (dissimilar) appearance and motion have high (low) similarity.

Figure 5: Two-step segmentation. The green regions indicate the foreground. The intermediate result (left), the foreground in the intermediate result (white) plus the shorter trajectories (dark gray) (center), and the final segmentation (right).

as

$$L_u(0) = \frac{1}{|S_u|} \sum_{i \in S_u} -\log(P^i_{\text{loc}}), \tag{5}$$

$$L_u(1) = \frac{1}{|S_u|} \sum_{i \in S_u} -\log(1 - P^i_{\text{loc}}). \tag{6}$$

2.3 Pairwise Potential

The pairwise potential measures the affinity between two neighboring trajectories based on appearance and motion similarity (Fig. 4). Unlike the standard Potts potential, we allow it to be positive or negative, expressing both attraction and repulsion. Because the traditional Potts potential only take positive values, it prefer a shorter boundary, making it harder to segment thin objects. By allowing both signs, we avoid this tendency.

Since two neighboring trajectories are adjacent in at least one frame and share at least one boundary, we derive the affinity from the strength of color edges and motion edges along the shared boundaries. Either of the two edges alone is not sufficient to accurately model the affinity between trajectories. The color edges are often on the desired object boundary, but they can also be on other boundaries inside the foreground or background. The motion edge is useless when foreground is static. In addition, optical flow tends to be inaccurate around foreground boundary. Therefore, we define the pairwise potential to be a weighted sum of the color affinity $\psi_{\text{c}}(i, j)$ and the motion affinity $\psi_{\text{m}}(i, j)$, such as $V_{u,v}(l_u, l_v) = w_{\text{c}}\psi_{\text{c}}(u, v) + w_{\text{m}}\psi_{\text{m}}(u, v)$ if $l_u \neq l_v$, and 0 otherwise. For the color edge, we use the state of the art structured edge detector [14]. For the motion edge, we took the same approach as [9] to compute the motion boundary via optical flow. Both the color and the motion edges take values in $[0, 1]$. We define the affinity between two adjacent trajectories u and v by $\psi_{\text{c}}(u, v) = 2\exp(-\eta d_{\text{c}}(u, v)) - 1$, where η is a fixed parameter. The distance $d_{\text{c}}(u, v)$ between the two trajectories computes the average of the edge responses along the shared boundary in each frame, then maximize it over the frames in which the two trajectories are adjacent. The affinity $\psi_{\text{m}}(u, v)$ is computed similarly.

2.4 Optimization

We propose a two-step optimization procedure to obtain the final segmentation. In the first step, only relatively long trajectories are segmented, while second step segments all the trajectories after re-computing the unary potentials. This is based on the observation that regions with ambiguous unary potentials are often part of short TSP trajectories, which range from a single to several frames. Also, we employ the recently introduced non-submodular optimization method [15], since the use of the repulsion pairwise potential introduces supermodular pairwise potentials to the energy, preventing the use of the standard graph cuts.

Two-step Segmentation. Instead of segmenting all the trajectories at once, segmentation is done in two steps. In the first step, trajectories whose lengths are shorter than a predetermined length do not participate in the optimization. Only longer trajectories, which cover most of the background and some parts of the foreground, are segmented. The segmented regions are than used to re-estimate the unary potentials. In the second step, the remaining trajectories are added to the MRF and unary and pairwise potentials are updated. Optimization of the new energy gives the final segmentation. Because most of the background are part of the long trajectories and thus segmented in the first step, the refined background appearance model is highly reliable. The foreground appearance model also benefits from the partially segmented foreground regions. The refined unary appearance model helps segment regions with ambiguous initial unary potentials and correct errors made in the first step. The intermediate and the final segmentation results are shown in Fig. 5.

Local submodular approximation. Recent advance in MRF optimization and inference in a graphical model makes it possible to optimize a general non submodular energy effectively [15, 16]. Here, I employ in each step of the optimization the Local Submodular Approximation with Trust Region Strategy (LSA-TR) introduced in [15] that showed impressive results on difficult non submodular optimization problems [15, 17]. Briefly, LSA-TR approximate supermodular terms in the energy by the Taylor expansion at the current solution and iteratively refine the approximation in the trust region framework proposed by the same authors [18]. The algorithm internally invokes maximum flow computation a number of times. As used in this work, it is extremely fast because the number of optimization variables is small (the number of the TSP trajectories).

3 Experimental Validation

We evaluated the proposed approach on a subset of SegTrack dataset [19] that consists of four videos in which one or multiple foreground objects are densely labeled. The proposed method was applied to the four videos. The results are evaluated by average intersection over union (IOU) scores. This measure is computed by calculating the intersection divided by the union of the segmented masks and the ground truth masks, averaged over all frames.

Segmentation Results. Table 1 shows the average IOU scores. They are compared against the state of the art [8][9][10]. The most recent work is [8], currently the top-performing method in the SegTrack benchmark. The inside-outside map we used in this work was proposed in

Table 1: Average IOU score.

Name	Proposed	[8]	[9]	[10]
Bmx	0.76	0.79	0.67	0.17
Hummingbird	0.68	0.75	0.52	0.37
Soldier	0.68	0.83	0.69	0.60
Girl	0.61	0.91	0.73	0.82

Figure 6: The qualitative results. One of the frames (left) and its segmentation (center). The segmentation of another frame is also shown (right). From top to bottom: *Bmx*, *Hummingbird*, *Solider*, *Girl*.

[9], which is one of the approaches based on the object proposal [20]. It was previously the top-performing method on the original SegTrack dataset [4], which only consisted of five videos, all of which contained a single foreground. As far as we are aware, only [8] evaluates on the new dataset after it was updated in [19]. Scores of the previous work are taken from [8]. Qualitative results are shown in Fig. 6. Videos showing the results are also provided as the supporting material.

Discussion. Out of the four videos, the best result was achieved for *Bmx* both quantitatively and qualitatively. Its average IOU score is very competitive to [8]. For *Bmx* and *Hummingbird*, which contain two foreground objects, our results outperform [9] and [10]. However, note that [9] and [10] assume only a single foreground object in a video. For *Soldier* and *Girl*, which contain a single foreground, [9] and [10] perform well. Our score for *Soldier* is worsened by the segmented shadows: since the shadows move with the soldier and the method is unsupervised, there is no way to distinguish the shadows from the moving foreground. The result for *Girl* leaves much to be desired. Since all the pixels in the same trajectory are assigned the same label, our scores are also worsened by errors in TSP segmentation. Visual inspection of the results shows that this is especially true for *Hummingbird* and *Girl*.

Limitations. Since our method depends on the quality of TSP segmentation, if TSP segmentation does not work effectively our method would not be effective as well. For example, TSP segmentation was not effective for *Cheetah* sequence in the SegTrack dataset, which exhibits very fast motion with more than 40 pixels-displacements. It is very difficult to track superpixels under such fast camera motion. Out of more than 20000 unique trajectories, only eleven are longer than five frames. Our method would not yield a good result on such videos.

4 Conclusion

In this paper, we propose a new approach to unsupervised video segmentation based on trajectories of temporarily coherent superpixels. We cast the segmentation problem as a trajectory labeling problem with novel potential functions, which we minimize using a two-step optimization method. The proposed method is evaluated on a subset of the standard SegTrack benchmark and it can be seen that the algorithm allows more robust segmentation of the foreground that is static in some frames.

Acknowledgment

This work was partially supported by JST CREST.

References

[1] J Revaud, Matthijs Douze, Cordelia Schmid, and H Jegou, "Event retrieval in large video collections with circulant temporal encoding," in *CVPR*, 2013.

[2] M Jain, J. V Gemert, H Jegou, Patrick Bouthemy, and Cees G. M Snoek, "Action localization with tubelets from motion," in *CVPR*, 2014.

[3] D Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid, "Category-specific video summarization," in *ECCV*, 2014.

[4] David Tsai, Matthew Flagg, Atsushi Nakazawa, and James M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Computer Vision*, vol. 100, no. 2, pp. 190–202, Dec. 2011.

[5] S Vijayanarasimhan and Kristen Grauman, "Active frame selection for label propagation in videos," in *ECCV*, 2012.

[6] Suyog D. Jain and Kristen Grauman, "Supervoxel-consistent foreground propagation in video," in *ECCV*, 2014.

[7] David Varas and Ferran Marques, "Region-based particle filter for video object segmentation," in *CVPR*, 2014.

[8] Alon Faktor and Michal Irani, "Video segmentation by non-local consensus voting," in *BMVC*, 2014.

[9] Anestis Papazoglou and Vittorio Ferrari, "Fast object segmentation in unconstrained video," in *ICCV*, 2013.

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

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

[12] Jason Chang, Donglai Wei, and John Fisher III, "A video representation using temporal superpixels," in *CVPR*, 2013.

[13] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. PAMI*, 2011.

[14] Piotr Dollar and C. Lawrence Zitnick, "Structured forests for fast edge detection," in *ICCV*, 2013.

[15] Lena Gorelick, Yuri Boykov, Olga Veksler, Ismail Ben Ayed, and Andrew Delong, "Submodularization for binary pairwise energies," in *CVPR*, 2014.

[16] Paul Swoboda, Bogdan Savchynskyy, Jorg H. Kappes, and Christoph Schnorr, "Partial optimality by pruning for MAP-inference with general graphical models," in *CVPR*, 2014.

[17] Claudia Nieuwenhuis, Eno Toeppe, Lena Gorelick, Olga Veksler, and Yuri Boykov, "Efficient squared curvature," in *CVPR*, 2014.

[18] Lena Gorelick, Frank R. Schmidt, and Yuri Boykov, "Fast trust region for segmentation," in *CVPR*, 2013.

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

[20] Ian Endres and Derek Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Trans. PAMI*, 2014.