VIDEO OBJECT SEGMENTATION AGGREGATION

Tianfei Zhou[1,2], *Yao Lu*[1,2], *Huijun Di*[1,2], *Jian Zhang*[3]

Beijing Laboratory of Intelligent Information Technology[1]
School of Computer Science, Beijing Institute of Technology[2]
Faculty of Engineering and Information Technology, University of Technology Sydney[3]

ABSTRACT

We present an approach for unsupervised object segmentation in unconstrained videos. Driven by the latest progress in this field, we argue that segmentation performance can be largely improved by aggregating the results generated by state-of-the-art algorithms. Initially, objects in individual frames are estimated through a per-frame aggregation procedure using majority voting. While this can predict relatively accurate object location, the initial estimation fails to cover the parts that are wrongly labeled by more than half of the algorithms. To address this, we build a holistic appearance model using non-local appearance cues by linear regression. Then, we integrate the appearance priors and spatio-temporal information into an energy minimization framework to refine the initial estimation. We evaluate our method on challenging benchmark videos and demonstrate that it outperforms state-of-the-art algorithms.

***Index Terms*—** Video object segmentation, data fusion, appearance model

1. INTRODUCTION

We propose an approach for unsupervised video object segmentation, which aims to automatically separate foreground objects from background in a video. This task is of great importance because determining accurate object boundaries in videos is crucial for video summarization and human-computer interaction. In a complex scene, this problem is challenging due to object photometric and geometric variations, motion blur and background clutter, etc. In recent literature, many sophisticated algorithms [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] have been proposed to address these difficulties, and have shown significant performance improvement on public benchmarks.

However, as shown in Tab. 1, each method can work well on different sequences, but none of them can perform better than others for all the cases. For instance, while [10] achieves excellent performance in *girl* and *monkeydog* sequences, it yields inferior performance in *birdfall* and *cheetah*. Furthermore, consider the segmentation results for the exemplar algorithms on a single frame illustrated in Fig. 1. We can see that within the frame, only parts of the original object appear in each segmentation result, and different results are able to capture different parts, thereby complementing each other. Thus, it will be beneficial to combine these methods so that they can collaboratively contribute to better results.

Fig. 1. (Best viewed in color) Segmentation results of the 5th frame in *girl* sequence by different methods. a) KeySeg [2], b) FastSeg [7], c) DAGSeg [6], d) SeamSeg [8], e) JOTSeg [9], f) Ours. We can see that none of the other 5 algorithms can perfectly delineate the contour of the girl. For instance, the segments in a) and b) are not accurate in terms of the girl's head, however, only in b), the two arms are segmented with excellent accuracy. Taking advantage of their complementary roles, we achieve more accurate segmentation result, as shown in f).

Upon these observations we suggest an aggregation method for video object segmentation that fuses multiple segmentation results into a better one. Our algorithm benefits from three main properties. First, within each frame, an initial coarse foreground estimation is effectively computed using majority voting, which accumulates all the segmentation maps in each frame generated by various existing methods. The maps represent regions likely to encompass different parts of a target object. Through fusion, we are able to exploit the complementary roles of these approaches and enable them to improve each other. Second, we train a long-term holistic

appearance model based on the initial foreground estimation using linear regression. The model helps improve the spatial accuracy of the segmentation as well as discover the missed parts in the initial estimation. Last, the initial result is refined by integrating the appearance and spatial-temporal information into an energy minimization framework.

To summarize, our main contributions are:

- an aggregation model that exploits the strengths of many state-of-the-art approaches for object segmentation in videos. To our knowledge, this is the first method that aims for improving segmentation quality with an aggregation model.
- a long-term holistic appearance model that are robust to partial occlusion, illumination variations, etc.

2. RELATED WORK

Early work on video object segmentation can be roughly divided into three categories: trajectory-based, superpixel-based and objectness-based.

Motion clustering of long-term point trajectories is a robust tool to extract moving objects from video shots, as recently demonstrated, *e.g.*, in [12, 3, 13, 14, 4]. The work formulates video segmentation as a spatio-temporal grouping problem in the trajectory domain and classifies the trajectories into foreground or background. The use of point trajectories in contrast to static appearance cues or per frame optical flow, provides temporally consistent clusters since grouping naturally propagates over time, and does not rely on motion information in a particular frame. However, the methods usually result in over-segmentation because low-level trajectories cannot accurately capture object-like appearance and motion.

A recent trend in video object segmentation is to exploit superpixel [7, 9, 10, 11] rather than pixelwise models. These approaches assume that an object can be formed by a combination of superpixels, and one can reliably extract it by grouping superpixels together according to spatio-temporal similarity [7, 11] or perceptual organization [10]. The use of superpixel provides a desirable computational reduction as well as video segmentation performance improvement.

Object proposal is one of the very-well researched area in computer vision, *e.g.*, [15, 16]. Their aim is to extract object likely regions that cover the entire object boundary with high accuracy. Recently, work of [2, 6, 5] has introduced object proposal into video object segmentation. These methods generally require two stages to segment objects in videos: object proposal generation, in which a pool of object-like segments are extracted as foreground candidates, and segment selection, which aims to select primary objects with accurate boundaries using both appearance and motion cues.

We observe from the above analysis that methods in different categories can only capture single aspect of objects, *e.g.*, pixelwise, superpixel-wise, or object-wise. Hence, we believe that one can largely improve the segmentation performance by fusing these methods with diverged properties. The idea has been employed to improve the performance of many vision applications, *e.g.*, visual saliency detection [17] and object tracking [18]. However, to our knowledge, the topic is novel in video object segmentation. In this work, we build on the aforementioned approaches and show how to aggregate them to obtain better segmentation quality.

3. VIDEO OBJECT SEGMENTATION AGGREGATION

In this section we present our approach to segment and track primary objects in videos. We begin by describing a baseline method to obtain an initial estimation. Then, we propose to refine the estimation in a spatio-temporal graph cut framework.

Given a video shot of F frames, we firstly compute a set of segmentation results $\{\mathcal{M}^i\}_{i=1:N}$ using N separate state-of-the-art video segmentation algorithms. The description for selection of the algorithms can be found in the Experiment section. Every result $\mathcal{M}^i$ is represented by F binary masks $\{\mathcal{M}_t^i\}_{t=1:F}$, where $\mathcal{M}_t^i$ indicates the segmentation result by the i-th algorithm at frame t.

3.1. The Baseline Method

Our baseline method attempts to combine binary segmentation masks in each frame using majority voting. Given the segmentation masks $\{\mathcal{M}^i\}_{i=1:N}$, we label the pixel p in each frame t as foreground or background by:

$$\mathcal{M}_t^*(p) = \begin{cases} 1 & \text{if} \quad \frac{1}{N}\sum_{i=1}^{N} \mathcal{M}_t^i(p) > 0.5, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Such a frame-independent method, although simple, achieves surprisingly accurate segmentation results, since it allows us to combine the strengths of various methods and meanwhile avoids their weakness. However, the baseline is limited in that 1) it cannot recover parts that are wrongly labelled by more than half of algorithms; 2) it estimates the label for each pixel individually without considering its spatial and temporal relationship with neighboring pixels.

3.2. Proposed Approach

In this section, we present our main aggregation model for video object segmentation. We address the limitations of the baseline method from two aspects: 1) An appearance model is trained by means of linear regression using information over the whole video. It can capture object parts wrongly segmented even by all the algorithms. 2) The initial estimation by the baseline is refined by integrating the appearance cues with a spatio-temporal graph cut framework, in which the segmentation task is formulated as a pixel labelling problem. Notice

that pixelwise segmentation is prohibitively computation expensive. Therefore, we opt for performing segmentation at the superpixel level.

Formally, given the superpixel set S_t at time t, in which each superpixel $s_t^i \in S_t$ is associated with a label $l_t^i \in \{0 : \text{background}, 1 : \text{foreground}\}$, our goal is to find, for all superpixels in all frames, an optimal labelling $\mathcal{L} = \{l_t^i\}_{t,i}$ according to the criterion:

$$\mathcal{L}^* = \underset{\mathcal{L}}{\operatorname{argmin}}\, E(\mathcal{L}) = \operatorname{argmin} \Bigg\{ \sum_{t,i} \Phi_g(l_t^i) + \sum_{t,i} \Phi_a(l_t^i) + \sum_{(i,j)\in\mathcal{E}_w} \Phi_w(l_t^i, l_t^j) + \sum_{(i,j)\in\mathcal{E}_b} \Phi_b(l_t^i, l_{t+1}^j) \Bigg\} \tag{2}$$

where the unary potentials $\Phi_g(l_t^i)$ and $\Phi_a(l_t^i)$ are measures of confidence for superpixel s_i belonging to an object according to the baseline method and appearance cues, respectively, while the pairwise potentials $\Phi_w(l_t^i, l_t^j)$ and $\Phi_b(l_t^i, l_{t+1}^j)$ respectively penalize the assignment of different labels to spatially or temporally similar neighboring superpixels.

3.2.1. *Mask Aggregation Model* Φ_g

The unary term Φ_g is calculated based on frame-level mask aggregation, as previously discussed in the baseline method. It defines the cost of assigning a superpixel as foreground according to the votes by a set of algorithms. Formally, for each superpixel s_t^i at time t, the unary term Φ_g is as follows:

$$\Phi_g(l_t^i) = \begin{cases} -\log(\phi_g(s_t^i)) & \text{if} \quad l_t^i = 1, \\ -\log(1 - \phi_g(s_t^i)) & \text{if} \quad l_t^i = 0 \end{cases} \tag{3}$$

where $\phi_g(s_t^i) = \sum_{p_t^{i,j} \in s_t^i} \mathcal{M}^*(p_t^{i,j})$ denotes the percentage of pixels $\{p_t^{i,j}\}_j$ in s_t^i that is voted by more than half of algorithms.

3.2.2. *Appearance Model* Φ_a

Appearance model is used to refine the initial segments which for complex objects are unlikely to be perfect. In each frame t, we extract two feature vectors from foreground region $\mathcal{M}_t^i$ and its surroundings, respectively. To obtain the surrounding region, we dilate the mask $\mathcal{M}_t^i$ using a flat disk-shaped structuring element with radius R ($R = 50$ in our experiments) to obtain a region that surely cover the entire object as well as its surroundings, as shown in Fig. 2(b). For simplicity, we use bag-of-words over RGB color features since color is demonstrated more robust than shape or motion features [19]. Given the features in all frames, we construct a few codewords, which are defined as centers of learnt clusters. In our task, it is hard to determine a cluster number that is suitable for all videos since they are often different in video length and frame resolution. Therefore, we use the Euclidean distance between features as a criterion to generate new centers. Given a feature, if its distances to all other centers are above a threshold T_1 ($T_1 = 10$ in our experiments) we regard it as the center of a new cluster; otherwise, the feature belongs to the existing nearest cluster.

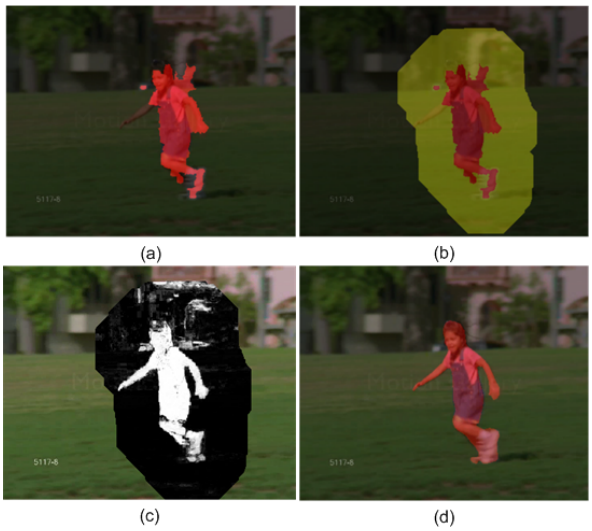

Fig. 2. (Best viewed in color) The illustration of our appearance model. (a) Voted result of frame #18 in *girl* sequence in Sec. 3.1. (b) Pixel set for the frame. Foreground pixels are shown in red, while background ones are shown in yellow. (c) The appearance prior computed by our appearance model. Note that the appearance model captures the missed parts of the target in the initial result, *e.g.*, the right arm of the girl. (d) The groundtruth.

Let $\mathbf{X} = [X_1|X_2|\dots|X_n]^\mathsf{T}$ denote a $n \times d$ feature matrix of a video where X_i is a 3-dimensional RGB color vector and $\mathbf{Y} = [Y_1|Y_2|\dots|Y_n]^\mathsf{T}$ be a binary indicator vector to indicate foreground or background pixels (*i.e.*, $Y_i = 1$ means X_i is a foreground pixel; $Y_i = 0$ means X_i belongs to background). Given a k-word codebook $\mathbf{C} = \{C_1, C_2, \dots, C_k\}$, we construct a $n \times k$ codebook histogram $\mathbf{Z}$ with each element calculated as $Z_{ij} = \exp(-\frac{\|X_i - C_j\|}{\tau})$, where $\tau = 20$ is a constant. Then, a regression weight matrix $\mathbf{W}$ is obtained by solving a least-square problem:

$$\min_{\mathbf{W}} \|\mathbf{Z}\mathbf{W} - \mathbf{Y}\|_F^2 + \lambda \|\mathbf{W}\|_F^2 \tag{4}$$

where $\|\cdot\|$ is the Frobenius norm. The optimal solution is given by the following system of linear equations:

$$(\mathbf{H} + \lambda \mathbf{I})\mathbf{W} = \mathbf{U} \tag{5}$$

where $\mathbf{H} = \mathbf{Z}^\mathsf{T}\mathbf{Z}$ is the covariance matrix, and $\mathbf{U} = \mathbf{Z}^\mathsf{T}\mathbf{Y}$ is the correlation matrix.

For a pixel i with color feature X_i, we firstly compute a codebook histogram $\mathbf{Z}_i = [Z_{i1}, Z_{i2}, \cdots, Z_{ik}]$ and thus the

probability of the pixel belonging to a target is equal to $\mathbf{Z}_i\mathbf{W}$. Then, the unary appearance term $\Phi_a(l_t^i)$ is defined as:

$$\Phi_a(l_t^i) = \begin{cases} -\log(\phi_a(s_t^i)) & \text{if} \quad l_t^i = 1, \\ -\log(1-\phi_a(s_t^i)) & \text{if} \quad l_t^i = 0 \end{cases} \tag{6}$$

where $\Phi_a(s_t^i)$ is the sum of priors of all pixels in superpixel s_t^i. Note that one can actually learn an appearance model for each algorithm and use them to vote the final appearance prior. However, we find that this does little help for performance improvement. Thus, we only train one single model based on the voted initial estimation.

3.2.3. *Pairwise Terms Φ_w and Φ_b*

The spatial and temporal smoothness terms encourage neighbouring superpixels with similar appearance to have the same label. In this work, we define these two energy terms following the conventional definitions in [7]. We refer readers to the corresponding paper for more details.

4. EXPERIMENTAL EVALUATION

In this section, we evaluate our video object segmentation method on the standard SegTrack dataset [1], which consists of 6 challenging videos (*birdfall*, *cheetah*, *girl*, *monkeydog*, *parachute* and *penguin*). All the videos are provided with pixel-level human annotated groundtruth for the primary foreground objects. We follow the setup in the literature [2] and use the first 5 videos for evaluation since the groundtruth for the *penguin* sequence is not usable.

4.1. Selection of Methods

Notice that the performance of basic algorithms used in our framework is significant for our method. Generally, if all the basic algorithms perform well in a sequence, we will also achieve good result. However, in many cases, the groundtruth is not available for some sequences, and hence it will be difficult to measure the performance of an algorithm. Therefore, we selected the state-of-the-art video object segmentation algorithms only according to the availability of public code. Moreover, we aimed to cover a large set of different working regimes: 1) one pixelwise method: [12], 2) two superpixel-based methods: [7] and [9], 3) two objectness-based methods: [2] and [6], 4) other method: [8].

Besides the selected methods, we also compare to other several approaches [1], [5], [11], [10], and our baseline method, as shown in Tab. 1.

4.2. Experimental Results

We use two metrics for quantitative analysis of the above methods. The first one is the average per-frame pixel error rate in comparison with groundtruth:

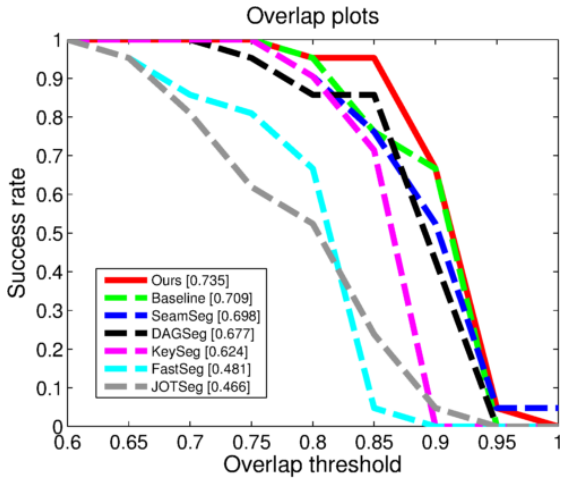

Fig. 3. Overlap plots for the SegTrack dataset. The methods are ranked using area under curve of each plot.

$$\text{err} = \frac{\text{xor}(\text{R}, \text{GT})}{F} \tag{7}$$

where R indicates the segmentation results of a video, and GT is the groundtruth of the video. F is the length of the video.

Furthermore, we introduce **overlap plot** to evaluate the overall performance of each method. The intersection-over-union overlap metric is used in [5] to overcome misleading of the first metric. However, we note that the metric still cannot measure the overall performance. Thus, in this paper, we count the number of successful frames whose overlaps are above a threshold T_2. The overlap plot shows the ratios of successful frames at the thresholds varied from 0 to 1. Using one success rate value at a specified threshold ($T_2 = 0.5$) may not be fair. Instead, we use the area under curve (AUC) of each overlap plot to rank the segmentation methods.

As can be seen from Tab. 1, we achieve comparable results using majority voting in our baseline method. Note that the baseline method is insensitive to the results of a single method. For example, although [12] performs poorly on most video sequences, our method is not affected by this approach and still produces satisfactory results. However, we also notice that on average, the baseline method still performs worse than the methods [8] and [10]. After refinement, we achieve great performance improvement on all sequences in comparison with the baseline. In particular, in three sequences (*birdfall*, *cheetah* and *parachute*), our method achieves the best segmentation performance in comparison with the state-of-the-art approaches.

Furthermore, we evaluate the overall performance of the methods, as illustrated in Fig. 3. The score for each method is calculated according to the area under curve of the corresponding overlap plot. We can clearly see that our method

Table 1. Results on the SegTrack dataset. We measure the average per-frame pixel error for each sequence. When compared with the state-of-the-art, our scheme outperforms all existing video object segmentation methods in terms of the average pixel error over the whole database.

	birdfall	cheetah	girl	monkeydog	parachute	average
Trajectory-based [12]	217	890	3859	284	855	868
SegTrack v1 [1]	252	1142	1304	563	235	594
KeySeg [2]	288	905	1785	521	201	592
FastSeg [7]	189	806	1698	472	221	542
DAGSeg [6]	155	633	1488	365	220	452
SegTrack v2 [5]	242	1156	1573	483	328	618
SeamSeg [8]	186	535	**761**	358	249	372
SaliencySeg [11]	209	796	1040	562	207	503
JOTSeg [9]	163	806	1904	342	275	528
SuperpixelSeg [10]	278	824	1029	**192**	251	397
Baseline	173	535	1408	316	215	414
Ours	**153**	**458**	1105	264	**180**	**342**

largely boosts the segmentation performance. Note that the rankings of some methods in Tab. 1 and Fig. 3 are slightly different. For instance, [8] is better in Tab. 1 than our baseline, while in Fig. 3 is worse. The reason for this lies in that in Tab. 1 we directly use the results of other methods (*e.g.*, [8]) presented in corresponding papers, while in Fig. 3 we re-ran their results using the authors' codes for overlap plots. The results we obtained are slightly different from those presented in the papers.

Fig. 4 illustrates the exemplar frames for the 5 video shots. We can clearly see that the proposed algorithm can accurately delineate the objects in most frames. In *birdfall* sequence, even though the bird is very small, our algorithm can localize the object and segment it with accurate boundaries. In *cheetah*, *girl* and *monkeydog*, we are able to obtain accurate boundaries of objects that are under large deformation. Although the results are slightly inaccurate due to motion blur and background clutter in frame #17 and #21 of *girl* sequence, our results are still satisfactory. In *parachute*, the target undergoes great illumination changes during falling down. Our method is insensitive to these changes and outperforms all other algorithms on this sequence.

5. CONCLUSION

This paper introduced a novel aggregation scheme for the challenging task of video object segmentation. In this scheme, an initial estimation is firstly computed according to a per-frame aggregation procedure using majority voting; a long-term holistic appearance model is then trained with linear regressor to refine the initial results; and the graph model encourages smoothness between superpixels that are spatio-temporally adjacent and similar in appearance. Both quantitative and qualitative results demonstrate that our method outperforms the state-of-the-art.

Acknowledgements: This work was supported in part by the National Natural Science Foundation of China (No. 61273273), by the Research Fund for the Doctoral Program of Higher Education of China (No. 20121101110034) and by the Specialized Fund for Joint Building Program of Beijing Municipal Education Commission.

6. REFERENCES

[1] Tsai David, Flagg Matthew, and M.Rehg James, "Motion coherent tracking with multi-label mrf optimization," in *BMVC*, 2010.

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

[3] Peter Ochs and Thomas Brox, "Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions," in *ICCV*, 2011.

[4] Katerina Fragkiadaki, Geng Zhang, and Jianbo Shi, "Video segmentation by tracing discontinuities in a trajectory embedding," in *CVPR*, 2012.

[5] Fuxin Li, Taeyoung Kim, A. Humayun, D. Tsai, and J.M. Rehg, "Video segmentation by tracking many figure-ground segments," in *ICCV*, 2013.

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

Fig. 4. From top to bottom, visual illustrations of our method over the sequences *birdfall*, *cheetah*, *girl*, *monkeydog* and *parachute*. Our method shows its ability to deal with different challenges: background clutter (*birdfall* and *cheetah*), object deformation (*cheetah*, *girl* and *monkeydog*) and illumination change (*parachute*).

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

[8] S. Avinash Ramakanth and R. Venkatesh Babu, "Seamseg: Video object segmentation using patch seams," in *CVPR*, 2014.

[9] Longyin Wen, Dawei Du, Zhen Lei, Stan Z. Li, and Ming-Hsuan Yang, "Jots: Joint online tracking and segmentation," in *CVPR*, June 2015.

[10] Daniela Giordano, Francesca Murabito, Simone Palazzo, and Concetto Spampinato, "Superpixel-based video object segmentation using perceptual organization and location prior," in *CVPR*, 2015.

[11] Wenguan Wang, Jianbing Shen, and Fatih Porikli, "Saliency-aware geodesic video object segmentation," in *CVPR*, 2015.

[12] Thomas Brox and Jitendra Malik, "Object segmentation by long term analysis of point trajectories.," in *ECCV*, 2010.

[13] Katerina Fragkiadaki and Jianbo Shi, "Detection free tracking: Exploiting motion and topology for segmenting and tracking under entanglement," in *CVPR*, 2011.

[14] Peter Ochs and Thomas Brox, "Higher order motion models and spectral clustering," in *CVPR*, 2012.

[15] Joao Carreira and Cristian Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *CVPR*, 2010.

[16] Ian Endres and Derek Hoiem, "Category independent object proposals," in *ECCV*, 2010.

[17] Long Mai, Yuzhen Niu, and Feng Liu, "Saliency aggregation: A data-driven approach," in *CVPR*, 2013.

[18] Christian Bailer, Alain Pagani, and Didier Stricker, "A superior tracking approach: Building a strong tracker through fusion," in *ECCV*, 2014.

[19] Horst Possegger, Thomas Mauthner, and Horst Bischof, "In defense of color-based model-free tracking," in *CVPR*, 2015.