Global Tracklet Matching for Unsupervised Video Object Segmentation

Xin Xiao, Changbin Cui, Yao Lu
Beijing Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, China
{xxiao, ccbingo, vis_yl}@bit.edu.cn

Abstract

*In this work, we consider a novel way to conduct object tracklet association in a holistic manner. Different from prior approaches that greedily address this problem, we attempt to find a global optimal matching of multiple tracklets using graph-based optimization. Specifically, our method is detailed as follows: we first generate instance segment proposals for each frame in a sequence, using a class-agnostic instance segmentation network. The proposals between consecutive frames are then locally matched using the Hungarian algorithm to obtain a collection of short-term tracklets. The greedy matching algorithm used here can only track the proposals with high similarity, and may fail due to occlusion or fast motion. Thus, it is necessary to further match the tracklets in order to obtain more meaningful tracking results. To this end, we formulate it as a graphcut problem over a tracklet graph, avoiding the expensive hierarchical clustering in the previous works. In particular, each node in the graph denotes a tracklet, while each edge measures the similarity between two tracklets. For each tracklet, we extract features for all the proposals belonging to it, and use the average feature as the final representation. To solve the graphcut clustering, we consider a fast, heuristic solution (*i.e.*, the Kernighan Lin (KL) algorithm), to partite the graph into different clusters, and each cluster can represent a segment track for an individual object. The proposed method ranked third place in the DAVIS-2020 Unsupervised Video Object Segmentation Challenge, on both test-dev and test-challenge benchmarks.*

1. Introduction

Unsupervised video object segmentation (UVOS) is a fundamental task to automatically segment and track primary objects in a video sequence without any prior information (*e.g.*, groundtruth of the first frame, or human annotations). It has a wide range of applications in video surveillance, intelligent robots, and autonomous driving.

Previous methods for UVOS mainly focus on single object segmentation. Most of them [11, 4, 7, 25, 26] use two-stream networks to extract appearance and motion features of objects and make use of these features to achieve the final segmentation results. Recently, unsupervised video multiple-object segmentation becomes a hot research topic, and many works [22, 14] are proposed for solving it. They first apply instance segmentation networks to generate instance proposals of each frame and then use a greedy matching algorithm to connect the proposals into multiple tracks. Although they achieve great performance, these methods often fail to track the objects due to occlusions or fast motion. We think the main reason is that the greedy matching method only utilizes the adjacent frames information rather than global information of the sequence to associate proposals, which leads to the sub-optimal results.

To solve this problem, we propose to associate proposals using global tracklet matching. Specifically, we first use a greedy matching method to associate proposals generated by the instance segmentation network into short tracklets. Then, we construct an undirected tracklet graph whose nodes are short tracklets and edges are the appearance similarity between tracklets. Finally, we formulate the segmentation task into a minimum cost subgraph multi-cut problem and apply a heuristic algorithm to get the feasible solutions as the final segmentation results. To evaluate the effectiveness of our method, we perform experiments on the DAVIS2020 challenge. Our method achieves the $\mathcal{J}\&\mathcal{F}$ Mean of 52.3 and the third place on the DAVIS2020 challenge.

2. Related Work

Instance Segmentation. Instance segmentation is a classical computer vision task which requires predicting the category and location of each object as well as its pixel-level segmentation mask. Previous methods can be roughly divided into one-stage or two-stage paradigms. One-stage methods are based on anchor-free object detection techniques [2, 24, 23] or instance embedding clustering [5]. These methods avoid manually designing achors, leading to more flexible and high efficient solutions. Two-stage approaches [6, 3] generally add a segmentation head on top of the conventional two-stage detection framekwork, achieving higher segmentation accuracy than one-stage methods. In our approach, we utilize the state-of-the-art instance segmentation technique, *i.e.*, Hybrid Task Cascade Network (HTC) [3], to generate instance segment proposals at each frame for a video. These proposals are considered as object candidates which will be temporally linked to obtain consistent tracks.

Multiple-object Tracking. Most current multiple-object tracking algorithms are based on the tracking by detection framework , which firstly gets the detection results of each frame and associates the detections into tracks. Among them, *subgraph decomposition* [19, 10] is an effective way to handle the data association problem. Specifically, they firstly build an undirected graph whose node are detections of each frame and whose edges are the spatial [19] or appearance similarity [10] of every two detections. After that, they formulate the undirected graph into a minimum cost subgraph multi-cut problem, and the feasible solutions to this problem are the final tracks. In this work, we solve a minimum cost subgraph multi-cut problem in UVOS.

Unsupervised Video Object Segmentation. Early UVOS methods typically locate objects in videos via object motion (*e.g.*, short-term optical flow [16], or long-term pixel trajectory [17]), or appearance (*e.g.*, objectness [1, 15] for object-like regions). Recently, many deep learning based methods are proposed for UVOS. One typical pipeline [18, 21, 11] is to extract appearance features by segmentation networks and exploit temporal consistency between appearance features to locate the primary objects. Another common approach is designing two-stream fully convolution networks [4, 20, 8] for extracting multi-modal features. Different streams encode either spatial or temporal information, and their fusion could provide a more comprehensive spatiotemporal object representation for more accurate segmentation. More recently, several approaches [22, 14] address unsupervised video multiple-object segmentation task. They employed a greedy matching algorithm to associate the instance masks between adjacent frames into final tracks. Different from [22, 14], we first generate short tracklets of each sequence and then apply a global tracklet matching algorithm to conduct object tracklet association in a holistic manner.

3. Proposed Method

As shown in Fig. 1, we propose a global tracklet matching method for UVOS. Our method mainly consists of three stages. The first stage is tracklet generation. We apply a class-agnostic instance segmentation network to process each video sequence to obtain the instance mask proposals and associate them into short tracklets depending on their spatial location. The second stage is tracklet graph construction. We build an undirected graph whose nodes are short tracklets and whose edges are the similarities of every two tracklets. The third stage is the subgraph decomposition. We formulate the tracklet graph as the minimum cost subgraph multi-cut problem and use a heuristic algorithm to get the feasible solutions to the problem, which are the final tracks we need. The implementation details are introduced as follows.

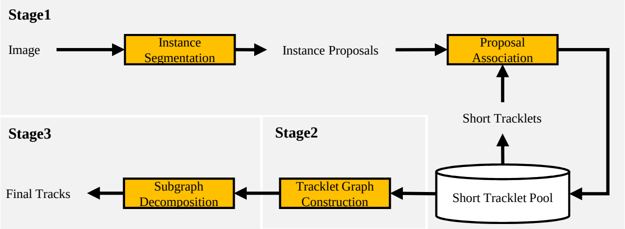

Figure 1: Overview of the global tracklet matching method. It mainly consists of three stages. Stage1: Tracklet Generation; Stage2: Tracklet Graph Construction; Stage3: Subgraph Decomposition.

Tracklet Generation. We first apply the Hybrid Task Cascade Network (HTC) [3], which has trained on COCO [12] to generate instance mask proposals and their corresponding detection scores. The version of ResNeXt-101-FPN based HTC is chosen as our implementation. Then, we construct bipartite graphs for adjacent frames of each sequence. The nodes of the bipartite graphs are the instance proposals of each frame, and the edges of the bipartite graphs are the spatial similarities of the two instance proposals. We obtain the spatial similarity by calculating the IoU between every two proposals. According to the IoU values, the Hungarian algorithm is used to associate each frame's proposal to obtain multiple short tracklets of each sequence. We compute the average detection score of all proposals in each tracklet as the corresponding tracklet score.

Tracklet Graph Construction. For each video sequence, we constructed an undirected tracklet graph $G = (V, E)$. The nodes V are short tracklets. The edges E are the appearance similarities of every two adjacent tracklets. For the calculation of appearance similarities, we first use a triplet based Re-ID network [13] to extract the appearance features for each proposal, and take the average of all the proposals in each tracklet as its representation. We use the Euclidean distance to measure the appearance similarity of every two tracklets.

Subgraph Decomposition. Following [19], we formulate the partition of the tracklet graph $G = (V, E)$ into a minimum cost subgraph multi-cut problem, which is summarized as follows:

$$\min_{\substack{x\in\{0,1\}^V \\ y\in\{0,1\}^E}} \sum_{v\in V} c_v x_v + \sum_{e\in E} d_e y_e, \tag{1}$$

$$\text{subject to} \quad \forall e = vw \in E: \quad y_{vw} \leq x_v, \tag{2}$$

$$\forall e = vw \in E: \quad y_{vw} \leq x_w, \tag{3}$$

$$\begin{aligned} &\forall C \in \text{cycles}(G)\ \forall e \in C: \\ &(1-y_e) \leq \textstyle\sum_{e'\in C\setminus\{e\}} (1-y_{e'}), \end{aligned} \tag{4}$$

where $c_v \in \mathbb{R}^V$ represents the unary features, and $d_e \in \mathbb{R}^E$ represents the pairwise features. We define the unary features as the tracklet scores and the pairwise features as the appearance similarities between each pair of tracklets. $G' = (V', E')$ denotes the feasible solutions of graph $G = (V, E)$. The solutions are encoded by $x \in \{0,1\}^V$ and $y \in \{0,1\}^E$, where the subset $V' = \{v \in V | x_v = 1\} \subseteq V$ of nodes and the subset $E' = \{vw \in E | y_{vw} = 1\} \subseteq E$ of edges. Eq. (2)) and Eq. (3) indicate that only nodes with two edges are selected at the same time the edge can be selected. Eq. (4) means that if one tracklet is connected to another tracklet, all neighbors of the first tracklet must be connected to all spatial and temporal neighbors of the second tracklet. We apply the Kernighan Lin (KL) algorithm [9] to acquire the feasible solution of G under the constrains (2), (3) and (4), which are the final segmentation results.

4. Experiments

We evaluate the performance of our method on DAVIS2020 challenge Unsupervised Video Object Segmentation Track (UVOS2020). The dataset of UVOS2020 consists of three parts: Train + Val unsupervised subset, Test-Dev 2019 subset (U19 T-D), and Test-Challenge 2019 subset (U19 T-C). We submit our final results to Codalab, and the results are presented as follows.

Table 1 reports the quantitative results on both U19 T-D and U19 T-C benchmarks. Compared with other competitors, our method achieves favorable results, *i.e.*, $\mathcal{J}\&\mathcal{F}$ Mean of 52.3 on U19 T-C and $\mathcal{J}\&\mathcal{F}$ Mean of 54.4 on U19 T-D. Our method performs well on the above two subsets, which demonstrates that our approach has high generalization abilities. The qualitative results on six sequences are shown in Fig. 2. We can see that our method performs well in cases of multiple challenging factors (*e.g.*, occlusions, fast motion, background clutter). This further verifies the effectiveness of our approach.

Table 1: Quantitative comparisons of different methods on the DAVIS 2020 Challenge. U19 T-D denotes the Test-Dev 2019 subset, and U19 T-C denotes Test-Challenge 2019 subset.

			TeamPhoenix	IIAI	BLIIT(Ours)	HCMUS
U19 T-D	$\mathcal{J}\&\mathcal{F}$	Mean ↑	57.9	**59.8**	54.4	-
	$\mathcal{J}$	Mean ↑	52.9	**56.0**	51.4	-
		Recall ↑	60.4	**65.1**	59.9	-
		Decay ↓	16.7	7.8	**-1.0**	-
	$\mathcal{F}$	Mean ↑	63.0	**63.7**	57.4	-
		Recall ↑	**69.5**	68.4	61.6	-
		Decay ↓	20.5	11.0	**0.0**	-
U19 T-C	$\mathcal{J}\&\mathcal{F}$	Mean ↑	**61.6**	55.6	52.3	43.9
	$\mathcal{J}$	Mean ↑	**58.4**	53.1	50.2	40.2
		Recall ↑	**65.0**	60.0	57.5	45.7
		Decay ↓	-1.6	-0.5	**-5.0**	-0.6
	$\mathcal{F}$	Mean ↑	**64.7**	58.2	54.4	47.5
		Recall ↑	**71.1**	62.5	58.9	50.1
		Decay ↓	0.5	1.6	**-2.5**	4.0

5. Conclusion

In this paper, we propose our global tracklet matching method for UVOS. We construct a tracklet graph with the global information of each sequence and formulate it into a minimum cost subgraph multi-cut problem whose feasible solutions are the final tracks that we need. Quantitative and qualitative results prove the effectiveness of our method, and we achieve the third place in the DAVIS2020 unsupervised video object segmentation challenge.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China (No.61273273), and by the National Key Research and Development Plan, China (No.2017YFC0112001).

References

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *CVPR*, 2014. 2

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: real-time instance segmentation. In *ICCV*, 2019. 1

[3] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 2

[4] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, 2017. 1, 2

[5] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 1

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[7] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017. 1

[8] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017. 2

[9] Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *The Bell system technical journal*, 49(2):291–307, 1970. 3

[10] Margret Keuper, Siyu Tang, Yu Zhongjie, Bjoern Andres,

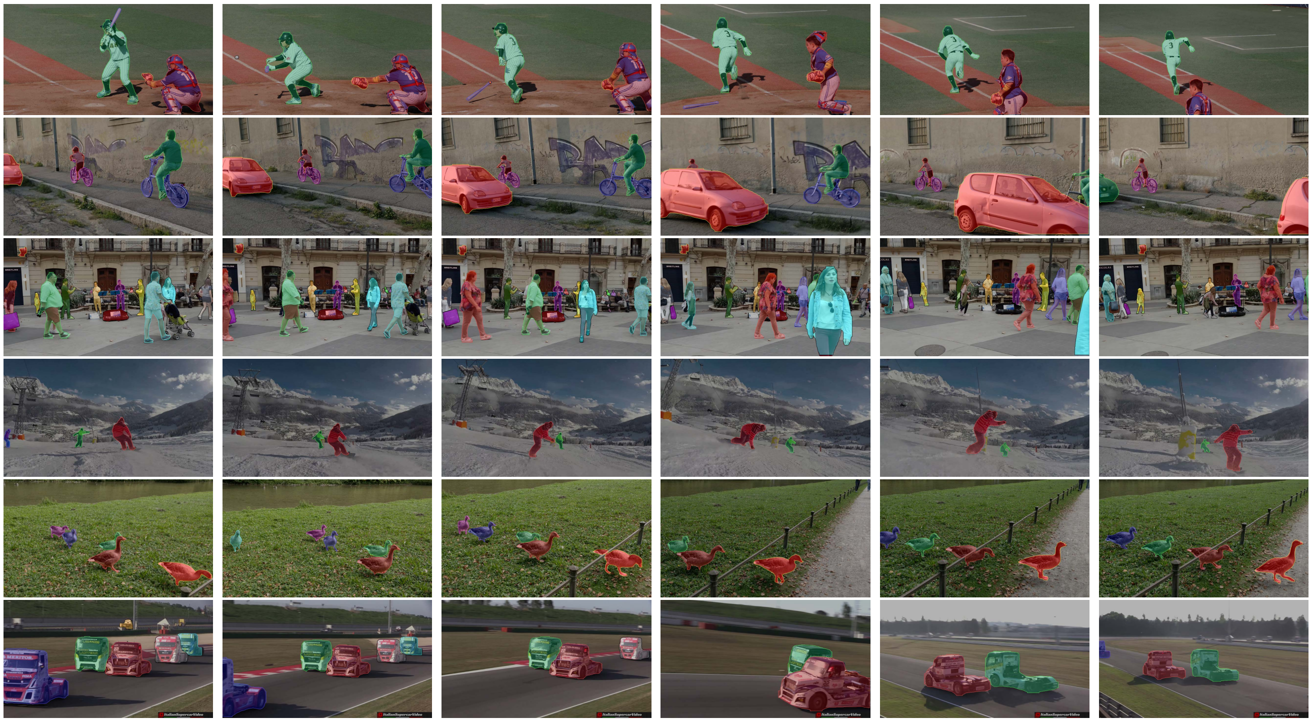

Figure 2: Qualitative results on DAVIS2019 test-dev and test-challenge. From top to bottom: *baseball*, *city-ride* and *music-band* from test-challenge, *snowboard-race*, *ducks* and *trucks* from test-dev.

Thomas Brox, and Bernt Schiele. A multi-cut formulation for joint segmentation and tracking of multiple objects. *arXiv preprint arXiv:1607.06317*, 2016. 2

[11] Siyang Li, Bryan Seybold, Alexey Vorobyov, Xuejing Lei, and C-C Jay Kuo. Unsupervised video object segmentation with motion-based bilateral networks. In *ECCV*, 2018. 1, 2

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[13] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 2

[14] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. Unovost: Unsupervised offline video object segmentation and tracking. In *WACV*, 2020. 1, 2

[15] Tianyang Ma and Longin Jan Latecki. Maximum weight cliques with mutex constraints for video object segmentation. In *CVPR*, 2012. 2

[16] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013. 2

[17] Jianbo Shi, Geng Zhang, and K. Fragkiadaki. Video segmentation by tracing discontinuities in a trajectory embedding. In *CVPR*, 2012. 2

[18] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kinman Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018. 2

[19] Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *CVPR*, 2015. 2

[20] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. In *ICCV*, 2017. 2

[21] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. Rvos: End-to-end recurrent network for video object segmentation. In *CVPR*, 2019. 2

[22] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*, 2019. 1, 2

[23] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. *arXiv preprint arXiv:1912.04488*, 2019. 1

[24] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *CVPR*, 2020. 1

[25] Kai Xu, Longyin Wen, Guorong Li, Liefeng Bo, and Qingming Huang. Spatiotemporal cnn for video object segmentation. In *CVPR*, 2019. 1

[26] Tianfei Zhou, Shunzhou Wang, Yi Zhou, Yazhou Yao, Jianwu Li, and Ling Shao. Motion-attentive transition for zero-shot video object segmentation. In *AAAI*, 2020. 1