Video Object Segmentation with Referring Expressions

Anna Khoreva[1(✉)], Anna Rohrbach[2], and Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
khoreva@mpi-inf.mpg.de
[2] University of California, Berkeley, USA

Abstract. Most semi-supervised video object segmentation methods rely on a pixel-accurate mask of a target object provided for the first video frame. However, obtaining a detailed mask is expensive and time-consuming. In this work we explore a more practical and natural way of identifying a target object by employing language referring expressions. Leveraging recent advances of language grounding models designed for images, we propose an approach to extend them to video data, ensuring temporally coherent predictions. To evaluate our approach we augment the popular video object segmentation benchmarks, DAVIS_{16} and DAVIS_{17}, with language descriptions of target objects. We show that our approach performs on par with the methods which have access to the object mask on DAVIS_{16} and is competitive to methods using scribbles on challenging DAVIS_{17}.

1 Introduction

Segmenting objects at pixel level provides a finer understanding of video and is relevant for many applications, e.g. augmented reality, video editing, rotoscoping, and summarisation. Ideally, one would like to obtain a pixel-accurate segmentation of objects in video with no human input during test time. However, the current state-of-the-art unsupervised video object segmentation methods [9] have troubles segmenting the target objects in videos containing multiple instances and cluttered backgrounds without any guidance from the user. Hence, many recent works [1,10] employ a semi-supervised approach, where a mask of the target object is manually annotated in the first frame and the task is to accurately segment the object in successive frames. Although this setting has proven to be successful, it can be prohibitive for many applications. It is tedious and time-consuming for the user to provide a pixel-accurate segmentation and usually takes more than a minute to annotate a single instance. To make video object segmentation more applicable in practice, instead of costly pixel-level masks [6,8] propose to employ point clicks or scribbles to specify the target object in the first frame. However, on small touchscreen devices, such as tablets or phones, providing precise clicks or drawing scribbles using fingers could be cumbersome and inconvenient for the user.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11132, pp. 7–12, 2019.
https://doi.org/10.1007/978-3-030-11018-5_2

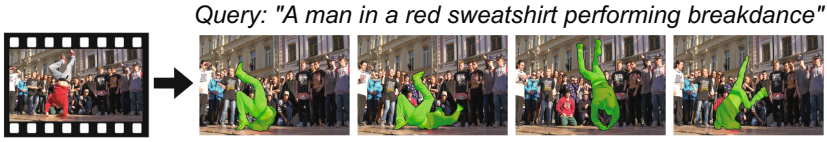

Fig. 1. Example result of the proposed approach.

To overcome these limitations, in this work we propose a novel task - segmenting objects in video using language referring expressions - which is a more natural way of human-computer interaction. It is much easier for the user to say: "I want the man in a red sweatshirt performing breakdance to be segmented" (see Fig. 1), than to provide a tedious pixel-level mask or struggle with drawing a scribble which does not straddle the object boundary. Moreover, employing language specifications can make the system more robust to background clutter, help to avoid drift and better adapt to the complex dynamics inherent to videos while not over-fitting to a particular view in the first frame.

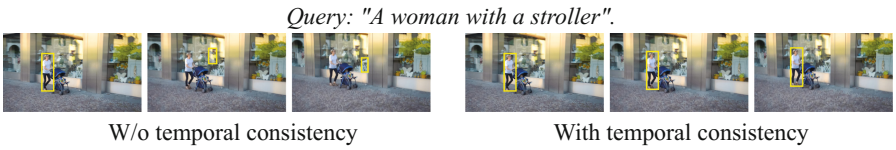

Fig. 2. Qualitative results of language grounding with and w/o temporal consistency.

We aim to investigate how far one can go while leveraging the advances in image-level language grounding and pixel-level segmentation in videos. We propose a convnet-based framework that allows to utilize referring expressions for video object segmentation, where the output of the grounding model (bounding box) is used as a guidance for segmentation of the target object in each video frame. To the best of our knowledge, this is the first approach to address video object segmentation via language specifications. For the extended version of this work we refer the reader to [5], collected language descriptions are available at https://www.mpi-inf.mpg.de/vos-language.

2 Method

Given a video $V = \{f_1, ..., f_N\}$ with N frames and a textual query of the target object Q, our aim is to obtain a pixel-level mask of the target object in every frame that it appears. Our method consists of two steps. Using as input the query Q, we first generate target object box proposals for every video frame by exploiting language grounding models, designed for images only. Applying these models off-the-shelf results in temporally inconsistent and jittery box predictions

(see Fig. 2). To mitigate this issue we next employ temporal consistency, which enforces boxes to be coherent across frames. As a second step, using as guidance the obtained box predictions of the target object on every frame we apply a segmentation convnet to recover detailed object masks.

Grounding Objects in Video. The task of language grounding is to localize a region described by a given language expression. It is typically formulated as measuring the compatibility between a set of object proposals $O = \{o_i\}_{i=1}^{M}$ and a given query Q. The grounding model provides as output a set of matching scores $S = \{s_i\}_{i=1}^{M}$ between a proposal and a query. The highest scoring proposal is selected as the predicted region.

We employ the state-of-the-art language grounding model – MattNet [11], to localize the object in each frame. However, using the grounding model designed for images and picking the highest scoring proposal for each frame lead to temporally incoherent results. Even with simple queries for adjacent frames that look very much alike, the model often outputs inconsistent predictions. To resolve this issue we propose to re-rank proposals by exploiting temporal structure along with the original matching scores. Since objects tend to move smoothly through space and in time, there should be little changes from frame to frame and the box proposals should have high overlap between neighboring frames. By finding temporally coherent tracks that are spread-out in time, we can focus on the predictions consistent throughout the video and give less emphasis to objects that appear for only a short period of time. The grounding model provides the likeliness of each box proposal to be the target object by outputting a matching score s_i. Then each box proposal is re-ranked based on its overlap with the proposals in other frames, the original objectness score and its matching score from the grounding model. Specifically, for each proposal we compute a new score: $\hat{s}_i = s_i * (\sum_{j=1,j\neq i}^{M} r_{ij} * d_j * s_j / t_{ij})$, where r_{ij} measures an IoU ratio between box proposals i and j, t_{ij} denotes the temporal distance between two proposals ($t_{ij} = |f_i - f_j|$) and d_j is the original objectness score. Then, in each frame we select the proposal with the highest new score. The new scoring rewards temporally coherent predictions which likely belong to the target object and form a spatio-temporal tube.

Pixel-Level Segmentation. We exploit bounding boxes from grounding as a guidance for the segmentation network. The bounding box is transformed into a binary image and concatenated with the RGB channels of the input image and optical flow magnitude, forming a 5-channel input for the network. Thus we ask the network to learn to refine the provided boxes into accurate masks. As our architecture we build upon [2].

We train the network on static images, employing the saliency segmentation dataset [3] with a diverse set of objects. The bounding box is obtained from the ground truth masks. To make the system robust during test time to sloppy boxes from the grounding model, we augment the ground truth box by randomly jittering its coordinates (uniformly, $\pm 20\%$ of the original box width and height).

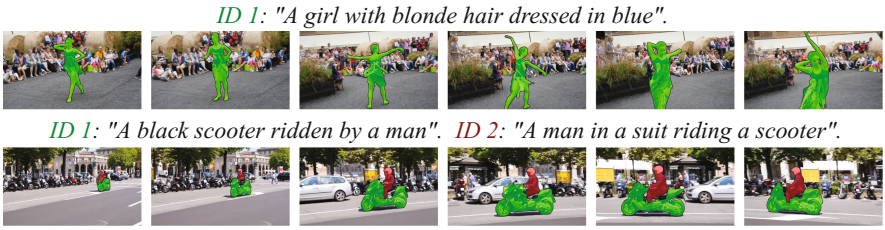

Fig. 3. Qualitative results using only referring expressions as supervision.

We synthesize optical flow from static images by applying affine transformations for both background and foreground object to simulate the camera and object motion in the neighboring frames. This simple strategy allows us to train on diverse set of static images, while exploiting motion information during test time. We train the network on many triplets of RGB images, synthesized flow magnitude images and loose boxes in order for the model generalize well to different localization quality of grounding boxes and different dynamics of the object. During inference we estimate optical flow with Flow-Net2.0 [4].

We make one single pass over the video, applying the model per-frame. The network does not keep a notion of the specific appearance of the object in contrast to [1], where the model is fine-tuned during the test time. Neither do we do an online adaptation as in [10], where the model is updated on its previous predictions. This makes the system more efficient during the inference and more suitable for real-world applications (Fig. 3).

3 Experimental Results

Here we present video object segmentation results using language referring expressions. To validate our approach we employ two popular datasets, DAVIS_{16} [7] and DAVIS_{17} [8], which we augmented with non-ambiguous referring expressions. We ask the annotator to provide a description of the object, which has a mask annotation, by looking only at the first video frame. For evaluation on DAVIS_{16} we use the mIoU measure and on DAVIS_{17} we employ $J\&F$ metric [8].

Table 1 compares our results to previous work. On DAVIS_{16} our method, while only exploiting language supervision, shows competitive performance, on par with techniques which use a pixel-level mask on the first frame (82.8 vs. 81.7 for OnAVOS [10]). Compared to [6] which uses click supervision, our method shows superior performance

Table 1. Results on $\text{DAVIS}_{16/17}$ validation sets.

Supervision	Method	DAVIS_{16} mIoU	DAVIS_{17} $J\&F$
1st frame mask	OSVOS [1]	80.2	57.0
	OnAVOS[a] [10]	81.7	59.4
Clicks	DEXTR [6]	80.9	-
Scribbles	Scribble-OSVOS [8]	-	39.9
Language	Our	82.8	39.3

[a]OnAVOS reports 86.1/67.8 on $\text{DAVIS}_{16/17}$ with online adaptation on successive frames.

(82.8 vs. 80.9). This shows that high quality results can be obtained via a more natural way of human-computer interaction – referring to an object via language, making video object segmentation more applicable in practice.

Lower numbers on DAVIS_{17} indicate that this dataset is much more difficult than DAVIS_{16}. Compared to mask supervision using language descriptions significantly under-performs. We believe that one of the main problems is a relatively unstable behavior of the underlying grounding model. There are a lot of identity switches, that are heavily penalized by the evaluation metric as every pixel should be assigned to one instance. The underlying choice of proposals for grounding could also have its effect. If the object is not detected, the grounding model has no chances to recover the correct instance. The method which exploits scribble supervision [8] performs on par with our approach. Note that even for scribble supervision the task remains difficult.

4 Conclusion

In this work we propose the task of video object segmentation using language referring expressions. We present an approach to address this new task as well as extend two well-known video object segmentation benchmarks with textual descriptions. Our experiments indicate that language alone can be successfully exploited to obtain high quality segmentations of objects in videos. We hope our results encourage further research on the proposed task and foster discovery of new techniques applicable in realistic settings, discarding tedious mask annotations.

References

1. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixe, L., Cremers, D., Gool, L.V.: One-shot video object segmentation. In: CVPR (2017)
2. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation (2017). arXiv:1706.05587
3. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. PAMI **37**, 569–582 (2015)
4. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. In: CVPR (2017)
5. Khoreva, A., Rohrbach, A., Schiele, B.: Video object segmentation with language referring expressions (2018). arXiv:1803.08006
6. Maninis, K., Caelles, S., Pont-Tuset, J., Gool, L.V.: Deep extreme cut: from extreme points to object segmentation. In: CVPR (2018)
7. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L.V., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)

8. Pont-Tuset, J., et al.: The 2018 Davis challenge on video object segmentation (2018). arXiv:1803.00557
9. Tokmakov, P., Alahari, K., Schmid, C.: Learning video object segmentation with visual memory. In: ICCV (2017)
10. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: BMVC (2017)
11. Yu, L., et al.: Mattnet: modular attention network for referring expression comprehension. In: CVPR (2018)