Video Object Linguistic Grounding

Alba Herrera-Palacio
albaherrerapalacio@gmail.com
Universitat Politecnica de Catalunya
Barcelona, Spain

Carles Ventura
cventuraroy@uoc.edu
Universitat Oberta de Catalunya
Barcelona, Spain

Xavier Giro-i-Nieto
xavier.giro@upc.edu
Universitat Politecnica de Catalunya
Barcelona, Spain

Figure 1: Example of the semi-supervised video object segmentation problem using language referring expressions from [3]

ABSTRACT

The goal of this work is segmenting on a video sequence the objects which are mentioned in a linguistic description of the scene. We have adapted an existing deep neural network that achieves state of the art performance in semi-supervised video object segmentation, to add a linguistic branch that would generate an attention map over the video frames, making the segmentation of the objects temporally consistent along the sequence.

CCS CONCEPTS

• **Computing methodologies** → **Video segmentation**; **Supervised learning**; *Natural language processing*; *Image segmentation.*

KEYWORDS

Video object gounding; neural networks; linguistics

ACM Reference Format:
Alba Herrera-Palacio, Carles Ventura, and Xavier Giro-i-Nieto. 2019. Video Object Linguistic Grounding. In *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications (MULEA '19), October 25, 2019, Nice, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3347450.3357662

1 INTRODUCTION

Video object segmentation (VOS) is a key step before performing any further processing in many applications. Semi-supervised VOS methods use a mask of the target object manually annotated in the first frame to accurately segment the object in successive frames. Results with semi-supervised methods seem to be successful, but the manual annotations are tedious and time-consuming.

MULEA '19, October 25, 2019, Nice, France

ACM ISBN 978-1-4503-6918-3/19/10…$15.00
https://doi.org/10.1145/3347450.3357662

The segmentation of target objects whose locations are roughly indicated by human interaction is referred to as interactive segmentation, where the costly pixel-level masks are replaced by point clicks [4] or scribbles [2] to specify the object in the first frame. A more natural way of human interaction is through referring expressions, textual queries, which can be applied to human-computer interactions. The goal of this work is a multimodal approach to segment target objects in video using language referring expressions.

Taking advantage from language grounding models that use referring expressions for image segmentation, we propose an approach where the output of the grounding model is used as a guidance for segmentation of the target object in each video frame. To evaluate our approach we use the video object segmentation benchmark DAVIS17 [5] with language descriptions of objects [3]. The next step of this project is to develop a fully end-to-end trainable model for video object grounding, where the language branch is not trained at the image level as done in [3], but at the video level.

2 STATE OF THE ART

2.1 Referring expression comprehension

Referring expression comprehension is the task of localizing an image region described by a natural language expression. Most existing approaches rely on external bounding box proposals which are scored to determine the top scoring box as the correct region. DBNet [10] is a discriminative bimodal neural network, where an expression and an image region serve as input and a binary classification decision is an output. MattNet [9] is a modular network which decomposes expressions into three modular components related to subject appearance, location, and relationship to other objects. Because of this, it is possible to easily adapt to expression containing different types of information.

2.2 Video object segmentation

VOS is a binary labeling problem aiming to separate foreground object(s) from the background region of a video.

Unsupervised methods assume no human input on the video during test time, also known as zero-shot segmentation. In this scenario, multiple object video object segmentation is a really hard

task, where the model must find the different objects to be segmented. They aim to group pixels consistent in both appearance and motion and extract the most salient spatio-temporal objects. Several techniques exploit object proposals, attention, such as [1], and optical flow. Since these methods do not have any knowledge about the object to track, they may have issues working on videos with multiple moving objects and cluttered backgrounds.

Semi-supervised methods assume human input for the first frame and then propagate the information to the successive frames. Most state-of-the-art semi-supervised methods rely on a pixel-accurate mask of an object provided for the first frame of a video, this is called one-shot video object segmentation. SiamMask [8] introduces a fully-convolutional siamese approach. Once trained, uses a single bounding box initialisation to produce class-agnostic object segmentation masks and rotated bounding boxes. FEELVOS [7] transfers information from the first frame and from the previous frame of the video to the current frame, using a semantic pixel-wise embedding together with a global and a local matching mechanism.

RVOS [6] proposes a fully end-to-end trainable network that incorporates recurrence on two different domains: the spatial to discover the different objects within a frame, and the temporal to keep the coherence of the segmented objects along time. RVOS has been trained for both zero-shot and one-shot video object segmentation.

2.3 Video language grounding

In this project we search for an alternative to identify an object by using language referring expressions. Apart from being more practical, language specifications allows to reduce the drift and make the system more robust.

Taking into account recent improvements in language grounding models designed for images, [3] suggests an approach to adapt them to video data. Given a referring expression, it first uses a grounding model for images to generate target object bounding box proposals for each frame. To mitigate temporally inconsistent predictions, it enforces temporal consistency, so that bounding boxes are coherent across frames. Finally, it uses the obtained box predictions of the target object to recover detailed object masks in each frame applying a convnet-based pixel-wise segmentation model. However, one of the drawbacks of this approach is that it is not trained end-to-end but depends on a network trained on static images.

3 METHOD

Given a video and several referring expressions of the target objects, we aim to obtain a pixel-level mask of those objects in every frame. To establish a baseline, a simple video object grounding method has been implemented using two pre-trained models: grounding model and semi-supervised video object segmentation model.

This method, as a first step, uses as input the first video frame and the target object textual query. It obtains a pixel-level mask of the target object in the first frame by exploiting a language grounding model designed for images only as MAttNet [9]. In multi-object segmentation, this first step is applied independently for each target object and the results are then fused in a color-coded pixel-level mask. Applying these off-the-shelf image-based models to each frame independently give temporally inconsistent results, so it is only used on the first frame.

Table 1: Results on DAVIS17 Test-dev.

Supervision	Method	*J&F*
First frame mask	RVOS [6]	50.3
Clicks	Scribble-OSVOS [2]	39.9
Language	Ours	21.9

As a second step, using the predicted pixel-level mask from the previous step and the video as inputs, the semi-supervised video object segmentation model RVOS[6] is applied to segment the objects found in the first frame by MattNet for the rest of the frames in the sequence. The temporal recurrence of the model enforces the temporal consistency of the segmentation results.

4 EXPERIMENTAL RESULTS

Here we present our video object segmentation results on the DAVIS17 dataset [5] with language referring expressions [3]. To validate our approach we employ *J&F* metric (see Table 1).

This method significantly under-performs compared to mask semi-supervised results due to the unstable behaviour of the pre-trained grounding model used. DAVIS17 is a multi-object segmentation dataset, in which more than one object is to be found in the first frame. Wrongly segmented objects by the grounding model cause occlusions and codifying all the objects in a color mask, where every pixel should be assigned to a single instance, leads to impossibility to recover the correct instances by the semi-supervised method. The evaluation metric heavily penalises those cases as the color code of some objects does not match the annotations provided (as it can be seen in Figure 2), affecting severely the results.

Figure 2: Example from golf sequence on DAVIS17. Left: ground truth. Right: MAttNet grounding results.

5 FUTURE WORK

In this work we have presented a basic video object grounding method using two pre-trained models: grounding model and semi-supervised video object segmentation model. The final goal of this project is to develop a fully end-to-end trainable model for multiple objects in video object segmentation using referring expressions.

We believe that in the future high quality results can be obtained referring to objects via language making video object segmentation more accessible and faster, and enhancing human-computer interaction.

ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (RTI2018-095232-B-C22 & TEC2016-75976-R).

REFERENCES

[1] Juan Leon Alcazar, Maria A Bravo, Ali K Thabet, Guillaume Jeanneret, Thomas Brox, Pablo Arbelaez, and Bernard Ghanem. 2019. MAIN: Multi-Attention Instance Network for Video Segmentation. *arXiv preprint arXiv:1904.05847* (2019).

[2] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. 2018. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557* (2018).

[3] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. 2018. Video object segmentation with language referring expressions. *arXiv preprint arXiv:1803.08006* (2018).

[4] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. 2018. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 616–625.

[5] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675* (2017).

[6] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marques, and Xavier Giro-i Nieto. 2019. RVOS: End-to-End Recurrent Network for Video Object Segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. 2019. FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation. *arXiv preprint arXiv:1902.09513* (2019).

[8] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2018. Fast Online Object Tracking and Segmentation: A Unifying Approach. *arXiv preprint arXiv:1812.05050* (2018).

[9] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *CVPR*.

[10] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. 2017. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 557–566.