The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Interactive Video Object Mask Annotation

Trung-Nghia Le[1]*, Tam V. Nguyen[2], Quoc-Cuong Tran[3, 5], Lam Nguyen[3, 5], Trung-Hieu Hoang[3, 5], Minh-Quan Le[3, 5], Minh-Triet Tran[3, 4, 5]

[1] National Institute of Informatics, Japan
[2] University of Dayton, U.S.A.
[3] University of Science, Ho Chi Minh City, Vietnam
[4] John von Neumann Institute, VNU-HCM, Vietnam
[5] Vietnam National University, Ho Chi Minh City, Vietnam

Abstract

In this paper, we introduce a practical system for interactive video object mask annotation, which can support multiple back-end methods. To demonstrate the generalization of our system, we introduce a novel approach for video object annotation. Our proposed system takes scribbles at a chosen key-frame from the end-users via a user-friendly interface and produces masks of corresponding objects at the key-frame via the Control-Point-based Scribbles-to-Mask (CPSM) module. The object masks at the key-frame are then propagated to other frames and refined through the Multi-Referenced Guided Segmentation (MRGS) module. Last but not least, the user can correct wrong segmentation at some frames, and the corrected mask is continuously propagated to other frames in the video via the MRGS to produce the object masks at all video frames.

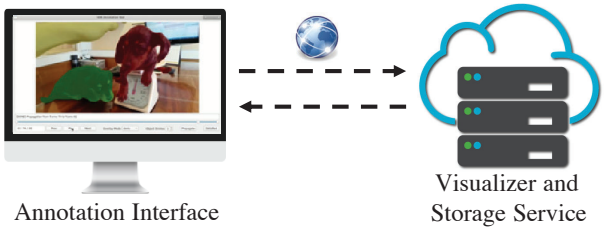

Figure 1: Our proposed annotation system, including Annotation Interface and Visualizer and Storage Service. Our system is flexible and supports multiple back-end methods for comparison and visualization.

Introduction

Deep learning-based video object detection/segmentation methods require a massive amount of training data (Le et al. 2020a; Le and Sugimoto 2019). However, manually labeling object masks for large-scale video datasets is a tedious and extremely time-consuming task. Hence, it is essential for the development of effective annotation frameworks which reduces human effort while maintaining high annotation accuracy. Indeed, efficient annotation tools are important and helpful in facilitating data annotation in multiple domains, *e.g.*, autonomous driving cars (Le et al. 2020b), drones (Zhu et al. 2020), wildlife preservation (Le et al. 2019), medical (Gelasca et al. 2008), and security (Ge et al. 2017).

Several systems have been developed for interactive object mask segmentation in images (Benenson, Popov, and Ferrari 2019; Zhang et al. 2020; Ling et al. 2019; Acuna et al. 2018). However, these systems are inappropriate to apply for video processing due to frame-by-frame processing. Few video instance segmentation methods have been introduced (Bertasius and Torresani 2020; Yang, Fan, and Xu 2019), but these methods only work on learned categories. These methods are unable to be used to annotate new categories. Semi-supervised video object segmentation methods have been proposed to segment video objects regardless their categories (Le et al. 2017; Tran et al. 2018, 2019; Oh et al. 2019a; J. Luiten 2018). These methods propagate object masks in key-frames with high accuracy and can be utilized for video object mask annotation.

In this work, we develop a practical system for interactive video object mask annotation. Our annotate system includes a local Annotation Interface and a Visualizer and Storage Service (as shown in Figure 1). Our system is flexible and can support multiple back-end methods for comparison and visualization.

To demonstrate the generalization of our system, we propose a novel interactive video object mask annotation approach for the Annotation Interface. Our system consists of a front-end interaction interface and two back-end key modules. We design a user-friendly interface (as shown in Figure 2), which allows the end-users to identify objects-of-interest, track, and correct objects in a feedback loop with the system. Input scribbles are transformed into masks of corresponding objects via the Control-Point-based Scribbles-to-Mask (CPSM) module and propagated to other video frames by Multi-Referenced Guided Segmentation (MRGS) module. Our system also allows the user to correct the wrong segmentation in certain frames, and the corrected masks are continuously propagated to other frames in the video via the MRGS module. Our annotation system will be public at our project page[1].

*E-mail address: *ltnghia@nii.ac.jp*

[1] https://sites.google.com/view/ltnghia/research/interactive-video-object-mask-annotation

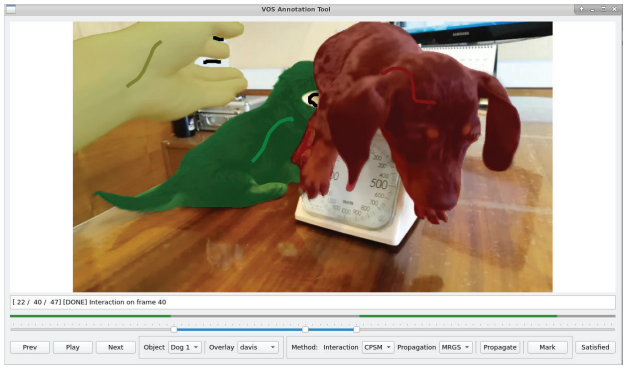

Figure 2: Our user-friendly interface for video object mask annotation.

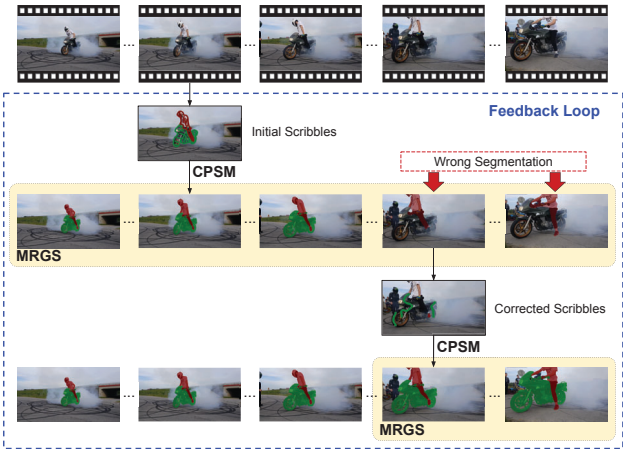

Figure 3: Overview of our interactive annotation flow. The users interact with the system via scribbles in a feedback loop. The system generates object masks from scribbles by CPSM module and propagate them over video frames via MRGS module. Meanwhile, the users correct object masks by new scribbles.

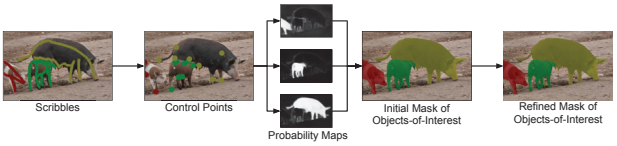

Figure 4: An example of Control-Point-based Scribbles-to-Mask. The figure is best view when magnified at 200%.

Proposed Annotation Approach

Interactive Annotation Flow

Figure 3 illustrates the interactive annotation flow of our proposed system, which consists of multiple rounds of interaction in a feedback loop. Our method has two stages for each round: interaction from the end-users to generate object masks at the chosen key-frame, and mask propagation at other frames.

In the first interaction round, the end-users input scribbles to create initial object masks at a selected key-frame through Control-Point-based Scribbles-to-Mask (CPSM) (Tran, Vu-Le, and Tran 2020). Then, Multi-Referenced Guided Segmentation (MRGS) is performed to produce masks of objects-of-interest at other frames (Tran et al. 2020). From the second interaction round onward, the end-users input scribbles to correct wrong segmentation at certain frames via CPSM. MRGS then propagates corrected masks to adjacent frames.

Control-Point-based Scribbles-to-Mask

We propose Control-Point-based Scribbles-to-Mask (CPSM) (Tran, Vu-Le, and Tran 2020) module to generate object masks from scribbles input of users. The input scribbles are first converted into control points using a *Control-points Extractor*. To extract control points from a scribble path, we evenly sample points along that path. Each of these control points holds the information about its coordinates on the image and the object to which it belongs. Then, control points are used by a *Control-points-to-Mask* module to generate a complete multi-object segmentation mask using a backpropagating refinement algorithm (Sofiiuk et al. 2020). We remark that if it is not the first interaction (*i.e.*, correcting wrong segmentation), a previously inferred mask is available and combined with this mask.

Finally, *Self-Reference Refinement* is performed to refine the generated mask, using memory-based model STM (Oh et al. 2019b). We develop a Global Memory Pool that is global across interactions to memorize all previously inferred masks. The memory-based model STM uses that memory pool to perform fine-grained segmentation. Figure 4 shows the flowchart from scribbles to masks.

Our CPSM module was combined with a simple object mask propagation algorithm and achieved second place on the Interactive Segmentation Track of the DAVIS 2020 Challenge with AUC 76.7 (Tran, Vu-Le, and Tran 2020).

Multi-Referenced Guided Segmentation

We propagate each object independently, then merge all object masks together. We propose Multi-Referenced Guided Segmentation (MRGS) (Tran et al. 2020) module by utilizing the structure of each object, together with its movement flow and deformable transformation, to create better segmentation for that object. In particular, the main steps in the process of each object are as follows.

First, we apply mask tracking and segmentation (Tran et al. 2019) to create an initial segmentation result of the object at video frames. Next, we propose *Single-Source Guided Segmentation* to quickly propagate an initial mask to all frames by performing fined-grained segmentation on a natural-shaped region-of-interest (ROI). Our method can reduce the ambiguity in the segmentation of different objects, especially those of the same category, in a regular rectangular region. We aim to transform a rectangular ROI to a

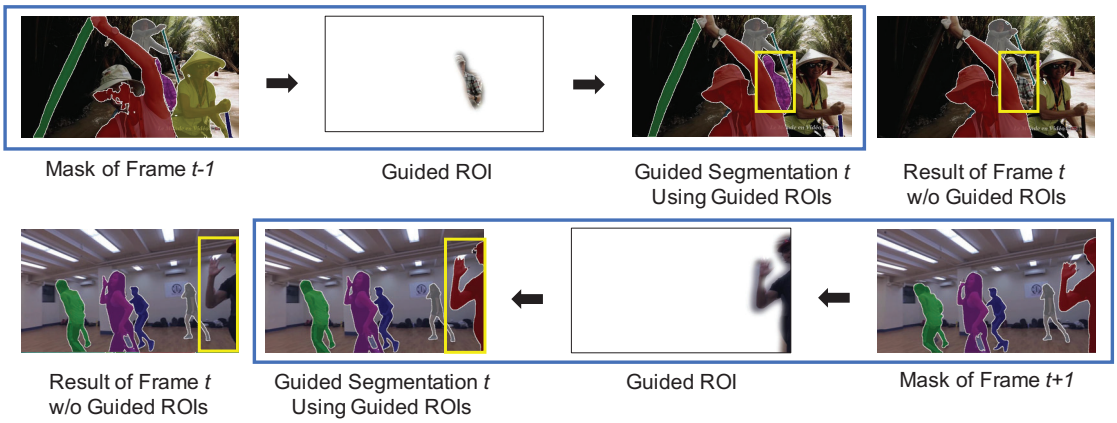

Figure 5: Bi-directional propagation strategy, including forward propagation and backward propagation, in our proposed Single-Source Guided Segmentation and Multi-Referenced Mask Propagation modules.

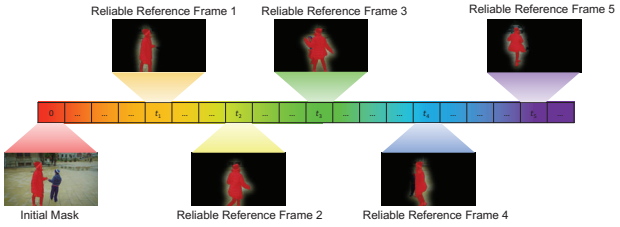

Figure 6: Our proposed Reliable Extra Samples module to collect reliable reference frames with corresponding guided natural-shaped ROI of the object-of-interest. The figure is best view when magnified at 200%.

non-rectangular ROI, namely natural-shaped ROI, across the object boundary to eliminate complex background inside the ROI. As discussed in (Nguyen, Zhao, and Yan 2018), attention cues are very useful for the segmentation task. Therefore, we propose a bi-directional strategy of guided attention to construct the natural-shaped ROI and then perform fine-grained segmentation on this guided ROI (c.f. Fig.5). Forward propagation strategy, where attention is referenced from initial segments of previous frames, can correct excessed segmentation due to dense objects in a ROI. Meanwhile, back-propagation strategy, where attention is referenced from initial segments of next frames, can recover missing objects due to fast motion, occlusion, or heavy deformation.

We then evaluate these results to select *Reliable Extra Samples* in the video for an object to create a set of reliable confidence reference frames for the object-of-interest (c.f. Fig. 6). Notably, we remove blurry frames, tiny objects, anomaly (i.e., sudden changes in object mask), and frames that are too similar (i.e., redundant frames).

Finally, we re-propagate masks with reference to multiple extra samples, namely *Multi-Referenced Mask Propagation*. We put all reliable reference frames in the memory pool for reference. We further propose to propagate masks from multiple reference frames instead of from a single-source. Each reference frame influences the results of the frames in its neighbors, and each frame usually depends on its nearest reference frame. To enhance the consistency of object masks across frames, we find anomaly in consecutive frames and correct them. An anomaly occurs when a mask accidentally disappears or re-appears in a short period, or its size and shape change significantly. In order to handle an anomaly case, we use both forward and backward mask propagation to restore missing segments and correct sudden changes in mask size and shape (c.f. Fig.5).

For fine-grained segmentation in guided natural-shaped ROIs, we train DeepLab3+ (Chen et al. 2018) model on MS-COCO (Lin et al. 2014) and DAVIS (Perazzi et al. 2016; Pont-Tuset et al. 2017) datasets.

Our MRGS module achieved fourth place on the Semi-Supervised Segmentation Track of the DAVIS 2020 Challenge with the global score of 79.3 (Tran et al. 2020).

Conclusion and Future Work

This paper presents a novel practical system for interactive video object mask annotation, which can generate dense per frame object masks with only scribbles on sparse key-frames provided by the end-users. Our system interacts with end-users through their scribbles via a user-friendly interface. Following the input scribbles, our proposed system maps the scribbles to object masks and then propagates the masks to segment the object-of-interest in all videos. In addition, the users can correct wrong segmentation via scribbles quickly. We believe that our proposed system will attract and significantly reduce the cost of video object mask annotation.

In the future, we plan to investigate more interaction between the users and the system. Furthermore, we aim to utilize different techniques for better segmentation.

Acknowledgements

This research is in part granted by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA19, National Science Foundation (NSF) under Grant No. 2025234, and JSPS KAKENHI Grant Number 20K23355.

References

Acuna, D.; Ling, H.; Kar, A.; and Fidler, S. 2018. Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++. In *CVPR*.

Benenson, R.; Popov, S.; and Ferrari, V. 2019. Large-Scale Interactive Object Segmentation With Human Annotators. In *CVPR*.

Bertasius, G.; and Torresani, L. 2020. Classifying, Segmenting, and Tracking Object Instances in Video with Mask Propagation. In *CVPR*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

Ge, S.; Li, J.; Ye, Q.; and Luo, Z. 2017. Detecting Masked Faces in the Wild with LLE-CNNs. In *CVPR*.

Gelasca, E. D.; Byun, J.; Obara, B.; and Manjunath, B. 2008. Evaluation and Benchmark for Biological Image Segmentation. In *ICIP*.

J. Luiten, P. Voigtlaender, B. L. 2018. PReMVOS: Proposal-generation, Refinement and Merging for the DAVIS Challenge on Video Object Segmentation. *CVPR Workshops* .

Le, T.; and Sugimoto, A. 2019. Semantic Instance Meets Salient Object: Study on Video Semantic Salient Instance Segmentation. In *WACV*, 1779–1788.

Le, T.-N.; Nguyen, K.-T.; Nguyen-Phan, M.-H.; Ton-That, V.; Nguyen, T.-A.; Trinh, X.-S.; Dinh, Q.-H.; Nguyen, V.-T.; Duong, A. D.; Sugimoto, A.; Nguyen, T. V.; and Tran, M.-T. 2017. Instance Re-Identification Flow for Video Object Segmentation. *CVPR Workshops* .

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch Network for Camouflaged Object Segmentation. *Journal of Computer Vision and Image Understanding* 184: 45–56.

Le, T.-N.; Sugimoto, A.; Ono, S.; and Kawasaki, H. 2020a. Attention R-CNN for Accident Detection. In *IEEE Intelligent Vehicles Symposium*.

Le, T.-N.; Sugimoto, A.; Ono, S.; and Kawasaki, H. 2020b. Toward Interactive Self-Annotation For Video Object Bounding Box: Recurrent Self-Learning And Hierarchical Annotation Based Framework. In *WACV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Ling, H.; Gao, J.; Kar, A.; Chen, W.; and Fidler, S. 2019. Fast Interactive Object Annotation with Curve-GCN. In *CVPR*.

Nguyen, T. V.; Zhao, Q.; and Yan, S. 2018. Attentive Systems: A Survey. *International Journal of Computer Vision* 126(1): 86–110.

Oh, S. W.; Lee, J.; Xu, N.; and Kim, S. J. 2019a. A Unified Model for Semi-supervised and Interactive Video Object Segmentation using Space-time Memory Networks. *CVPR Workshops* .

Oh, S. W.; Lee, J.-Y.; Xu, N.; and Kim, S. J. 2019b. Video object segmentation using space-time memory networks. In *ICCV*.

Perazzi, F.; Pont-Tuset, J.; McWilliams, B.; Gool, L. V.; Gross, M.; and Sorkine-Hornung, A. 2016. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *CVPR*.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Van Gool, L. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv:1704.00675* .

Sofiiuk, K.; Petrov, I.; Barinova, O.; and Konushin, A. 2020. f-BRS: Rethinking Backpropagating Refinement for Interactive Segmentation. *CVPR* .

Tran, M.-T.; Hoang, T.; Nguyen, T. V.; Le, T.-N.; Nguyen, E.; Le, M.; Nguyen-Dinh, H.; Hoang, X.; and Do, M. N. 2020. Multi-Referenced Guided Instance Segmentation Framework for Semi-supervised Video Instance Segmentation .

Tran, M.-T.; Le, T.-N.; Nguyen, T. V.; Ton-That, V.; Hoang, T.-H.; Bui, N.-M.; Do, T.-L.; Luong, Q.-A.; Nguyen, V.-T.; Duong, D. A.; and Do, M. N. 2019. Guided Instance Segmentation Framework for Semi-Supervised Video Instance Segmentation. *CVPR Workshops* .

Tran, M.-T.; Ton-That, V.; Le, T.-N.; Nguyen, K.-T.; Ninh, T. V.; Le, T.-K.; Nguyen, V.-T.; Nguyen, T. V.; and Do, M. N. 2018. Context-based Instance Segmentation in Video Sequences. *CVPR Workshops* .

Tran, Q.-C.; Vu-Le, T.-A.; and Tran, M.-T. 2020. Interactive Video Object Segmentation with Multiple Reference Views, Self Refinement, and Guided Mask Propagation. In *CVPR Workshops*.

Yang, L.; Fan, Y.; and Xu, N. 2019. Video instance segmentation. In *ICCV*, 5188–5197.

Zhang, S.; Liew, J. H.; Wei, Y.; Wei, S.; and Zhao, Y. 2020. Interactive Object Segmentation With Inside-Outside Guidance. In *CVPR*.

Zhu, P.; Wen, L.; Du, D.; Bian, X.; Hu, Q.; and Ling, H. 2020. Vision Meets Drones: Past, Present and Future. In *arXiv preprint arXiv:2001.06303*.