The 2018 DAVIS Challenge on Video Object Segmentation - CVPR 2018 Workshops

Context-based Instance Segmentation in Video Sequences

Minh-Triet Tran *[1], Vinh Ton-That[1], Trung-Nghia Le[3], Khac-Tuan Nguyen[1], Tu V. Ninh[1], Tu-Khiem Le[1], Vinh-Tiep Nguyen[2], Tam V. Nguyen[4], and Minh N. Do[5]

[1]University of Science, VNU-HCM, Vietnam
[2]University of Information Technology, VNU-HCM, Vietnam
[3]The Graduate University for Advanced Studies (SOKENDAI), Japan
[4]University of Dayton, US
[5]University of Illinois at Urbana-Champaign, US

Abstract

In this work, we propose Context-based Instance Segmentation for video object segmentation in two passes. Namely, in the first pass, we estimate the main properties of each instance (i.e., human/non-human, rigid/deformable, known/unknown category) by propagating its initial mask to other frames. We employ Instance Re-Identification Flow in this pass. The result of the first pass helps our system to automatically select the appropriate scheme for instance segmentation in the second pass. In the second pass, we process human and non-human instances separately. For human instance, we employ Mask R-CNN to extract human segments, OpenPose to merge fragments (in a frame), and object flow to correct and refine the result across frames. For non-human instance, if the instance has a wide variation in its appearance and it belongs to known categories (which can be inferred from the initial mask), we use Mask-RCNN for instance segmentation. If the instance is nearly rigid, we synthesize images from the first frame of a video sequence. We use affine and non-rigid deformations, together with illumination changes, to generate variants of the initial mask. To choose appropriate background for synthesized images, we retrieve images from the Places365 dataset having the similar scene category and scene attributes with the original frame. FCNs, including DeepLab2 and OSVOS, are trained on our synthesized dataset for each instance. For a deformable object in an unknown category, we reuse the baseline result from the first pass. Finally, instances in each frame are merged based on their depth values, using DCNF-FCSP, together with human and non-human object interaction and rare object priority.

*Corresponding author. Email: tmtriet@fit.hcmus.edu.vn

1. Introduction

Video object segmentation aims to label each video frame pixel to foreground instance object(s) or the background region. Recently, a new dataset [10] is constructed for DAVIS Challenge on Video Object Segmentation. This dataset is challenging due to multiple object instances with more distractors, *i.e.*, smaller instances and fine structures, more occlusions and fast motion.

There exists two main streams of approaches to solve this interesting problem. The first solution type is to perform instance re-identification [8, 7]. Each instance is detected and re-identified through frames. The second solution type is to perform instance segmentation with data augmentation [6]. The given first frame with ground truth is augmented with Lucid Data Dreaming [6] to generate more training data with different viewpoints. More discussions about recent works on this task can be found in [10].

We propose the Context-based Instance Segmentation (CIS) framework for video semantic segmentation. There are different schemes in CIS for instance segmentation adapting to the context, *i.e.* the category and visual properties of an instance. For human instances, we use OpenPose to control and refine the results from Mask-RCNN. For non-human instances, we train FCNs, including DeepLabs and OSVOS, from our synthesized dataset for nearly-rigid instances with similar background scenes. Mask-RCNN is used for deformable non-human instances in known categories. We use IRIF [7] as the baseline scheme.

To select the appropriate scheme, we employ our method IRIF [7] to propagate the initial mask of an instance to other frames and analyze the visual properties of segmented regions. Our results on DAVIS 2018 Challenge dataset highly indicate that our method is competitive among the top performing submissions.

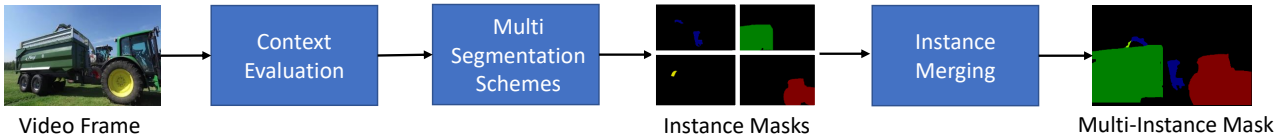

Figure 1. Overview of Context-based Instance Segmentation (CIS) framework.

Figure 2. Skeleton-guided region for unusual pose.

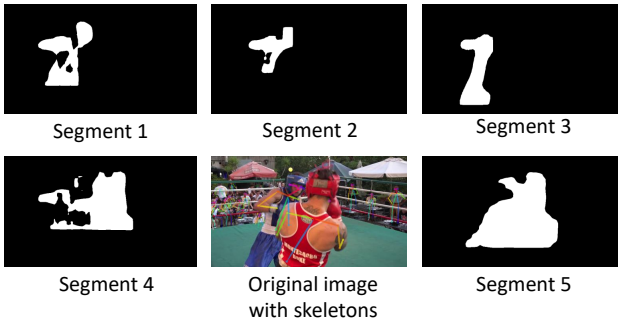

Figure 3. Human segments selection and linkage with reference to skeleton-guided regions.

2. Proposed Method

2.1. Overview

The key idea of our proposed method CIS is to determine the context of an instance so that we can apply an appropriate segmentation scheme for that instance. The context can be any observable properties that may affect the strategy to efficiently extract the mask of an instance in frames. Currently, we consider the following three attributes of an instance as its context: human or non-human, rigid or deformable, known or unknown category.

Figure 1 illustrates CIS with two phases: Context Evaluation and Context-based Segmentation. In the first phase, the category of an instance, such as person, car, dog, etc, can be directly inferred from its initial mask using Mask R-CNN [4]. To evaluate if an instance is rigid or deformable, we analyze the preview sequence of instance masks in the first $n_{Preview}$ frames. If there exists an homography matrix to transform the instance from the first frame to another frame for most frames in the first $n_{Preview}$ frames, we consider the instance to be rigid. To generate the preview mask sequence, we employ our method IRIF [7]. In the second phase, we propose multiple schemes corresponding to different context types: human instance, rigid non-human instance, deformable non-human instance in a known or unknown category. Each instance is processed independently over frames of a video sequence, and instance masks are blended with reference to depth information, human and non-human instance interaction, and rare instance priority.

2.2. Human Instance Segmentation

For *human instance*, we employ Mask-RCNN [4], pre-trained on the MS-COCO dataset , to extract human segments. As the results of Mask-RCNN may be affected by occlusion or unusual human pose, we use the skeletons from OpenPose [2] for reference to control and refine human instance segmentation.

For a human instance with an unusual pose which cannot be recognized by Mask-RCNN, we dilate the skeleton to obtain a *skeleton-guided region*, an image with only the region containing the complete human instance. We then apply Mask-RCNN on a skeleton-guided region. By eliminating unrelated content, Mask-RCNN has a higher chance to extract human instance segment correctly (Figure 2).

We also use the skeleton-guided region to select the human segments generated by Mask-RCNN that fit the human pose. In Figure 3, segment 4 does not well fit with any of the two skeleton-guided regions and should be discarded. The three segments 1, 2, and 3 can be linked as they all fit the same skeleton-guided region of the blue player. To preserve the inter-frame mask consistency, we use object flow [11] to correct and refine the result across frames.

2.3. Rigid Non-Human Instance Segmentation

For a rigid non-human instance, our objective is to accurately extract such instance from different backgrounds in the same scene category with the initial frame. Our method to process each instance is as follows. First, we synthesize images from the first frame of a video sequence. To generate visual variations of the initial mask, we apply both affine and non-rigid deformations, together with illumination changes, on the mask. We retrieve images from the Places365 dataset in the same scene category with the original background to preserve the semantic of the image. We train FCNs, including DeepLab2 [3] and OSVOS [1], on our synthesized dataset, namely wonderland data.

Wonderland Data Generation

Figure 4 (a) illustrates our proposed wonderland data generation. In this work, from a pair of input image and

mask, we generate $10,000$ different pairs of synthesized image and mask. The wonderland data is published on our website[1].

We collect scene photos from training set of the Places365 dataset. We manually discard artificial scenes, use only 22 natural scene categories with 592k images. For each image, we extract a feature at the last layer of DenseNet-161 [5], which was pre-trained on the Places365 dataset. This feature is used to build a hierarchical k-mean search (cf. Fig. 4 (b)) for each category independently.

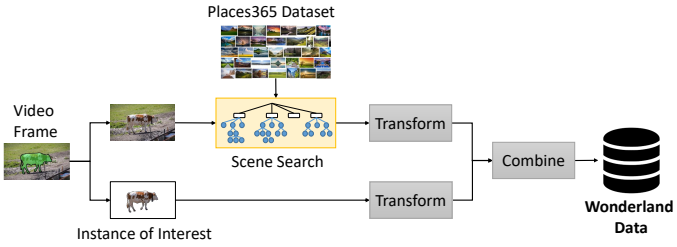

Figure 4. Wonderland data generation.

For an input image, we classify it into the corresponding category, using the pre-trained DenseNet-161 on the Places365 challenge dataset. After that, we search leaf nodes by comparing Euclidean distance between the feature of input image and the center of clusters.

We also extract object mask from the input image and transform the object and searched scenes independently, similarly to [6].

Training FCN

Offline Training: From pre-trained networks, including DeepLab2 [3] pre-trained on COCO-Stuff dataset and OSVOS [1] pre-trained on ImageNet dataset, we fine-tune these networks on the DAVIS training data. We remark that we use only the first frame of videos and apply the proposed wonderland data generation method for these images. These networks are considered as based networks.

Online Training: In the testing phase, for each video, we fine-tune based networks on the ground-truth mask of each instance at the first frame, called adaptive networks. Note that we also used wonderland data generated for each instance at the first video frame.

2.4. Deformable Non-Human Instance Segmentation in Known/Unknown Categories

For the non-human instances, we categorize them into two groups, namely, known and unknown categories. For the known categories, *i.e.*, already listed in MS-COCO dataset, we simply adopt Mask R-CNN to retrieve the instance segments. For the unknown categories, we obtain the segments via [7] since it can handle arbitrary object categories.

[1] https://sites.google.com/view/ltnghia/research/vos

2.5. Instance Merging

It is essential to determine the topology relationship (in term of z-order) between multiple instances so that we can sequentially combine corresponding masks of different instances into final result. We here merge instances based on depth values, human and non-human instance interaction, and rare instance priority heuristics as follows:

- **Depth values**: We first estimate pixel-wise depth values of the video frame, using DCNF-FCSP [9], and then take the average value for each instance.
- **Human and non-human instance interaction**: We define interaction heuristics as follow: transportation instances (such as horse, bike, motor, surfboard, and skateboard, etc.) are the farthest from the camera; human instance have the middle distance to the camera; and small non-human instances which can be hold, bring, touch, etc. are the nearest from the camera. Interacted small non-human instances are localized at human hand's position using OpenPose [2].
- **Rare instance priority**: Rare instances are identified and recovered similarly to [7]. We notice that rare instances are always the nearest ones from the camera.

3. Evaluation on 2018 DAVIS Challenge Dataset

As shown in Table 1, our CIS method achieves promising results in DAVIS Challenge 2018, namely, 64.1%, 68.6%, and 66.3% in terms of region similarity (Jaccard index), contour accuracy (F-measure), and global score, respectively. These results highly indicate that our method is competitive among the state-of-the-art methods in the DAVIS Challenge dataset. Our method maintains the performance as frames evolve as can be seen via the best performance in terms of J and F decay. We also note that the results of our proposed CIS method are better that those of using only IRIF [7] in the first pass, which are 61.5%, 6.62%, and 63.8%, respectively.

Furthermore, Figure 5 visualizes our video object segmentation results. From top row to bottom row, we can observe the first video frame, and a pair of processed video frames (IRIF [7] and our CIS, respectively). Our final CIS results successfully track and segment the key instances. Our framework can even handle the small instances, for example, the '*sphere*' in the '*juggle*' sequence (the second column from the left). More visual results can be found from our website[2].

4. Conclusion

In this paper, we introduce the novel CIS framework for multiple instances segmentation in videos. In particular, we

[2] https://www.selab.hcmus.edu.vn/research/vos/

Table 1. The performance of different methods in DAVIS Challenge 2018. The rankings in each categories are placed in parentheses. Our results are marked in **blue**.

#	Team	Global	Region J			Boundary F		
		Mean ⇑	Mean ⇑	Recall ⇑	Decay ⇓	Mean ⇑	Recall ⇑	Decay ⇓
1	Jono	**74.7** **(1)**	71.0 (2)	**79.5** **(1)**	19.0 (3)	**78.4** **(1)**	**86.7** **(1)**	20.8 (4)
2	Lixx	73.8 (2)	**71.9** **(1)**	79.4 (2)	19.8 (4)	75.8 (2)	83.0 (2)	20.3 (3)
3	Dawnsix	69.7 (3)	66.9 (4)	74.1 (5)	23.1 (5)	72.5 (3)	80.3 (5)	25.9 (5)
4	TeamILC_RIL	69.5 (4)	67.5 (3)	77.0 (3)	15.0 (2)	71.5 (4)	82.2 (3)	18.5 (2)
5	Apata	67.8 (5)	65.1 (5)	72.5 (6)	27.7 (9)	70.6 (5)	79.8 (6)	30.2 (9)
6	HCMUS	66.3 (6)	64.1 (6)	75.0 (4)	**11.7** **(1)**	68.6 (6)	80.7 (4)	**13.5** **(1)**
7	Alextheengineer	60.6 (7)	58.4 (7)	65.6 (7)	26.2 (7)	62.9 (7)	71.0 (8)	29.7 (8)
8	TeamVia	60.1 (8)	57.7 (8)	64.9 (8)	27.2 (8)	62.4 (8)	71.7 (7)	28.1 (6)
9	Kthac	58.9 (9)	56.7 (9)	63.1 (9)	30.7 (10)	61.1 (9)	67.6 (9)	33.1 (10)
10	TeamHuber99	54.5 (10)	51.8 (10)	56.4 (10)	25.7 (6)	57.2 (10)	62.5 (10)	29.5 (7)

Figure 5. The visualization results on the DAVIS Challenge 2018. From top to bottom: the first video frame with the ground-truth label followed by results of IRIF [7] and our CIS. The ground-truth of the certain video frame is not publicly available. Our final results significantly track and segment the instances of interest as annotated in the first frame.

incorporate more contextual information in order to handle rigid/non-rigid instances in known/unknown object categories. Throughout the experiments, our proposed framework surpasses our previous performance and achieves a competitive result among the leading submissions.

References

[1] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *CVPR*, 2017.

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, April 2018.

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988, 2017.

[5] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[6] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

[7] T.-N. Le, K.-T. Nguyen, M.-H. Nguyen-Phan, T.-V. Ton, T.-A. Nguyen, X.-S. Trinh, Q.-H. Dinh, V.-T. Nguyen, A.-D. Duong, A. Sugimoto, T. V. Nguyen, and M.-T. Tran. Instance re-identification flow for video object segmentation. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

[8] X. Li, Y. Qi, Z. Wang, K. Chen, Z. Liu, J. Shi, P. Luo, C. C. Loy, and X. Tang. Video object segmentation with re-identification. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

[9] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, June 2015.

[10] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[11] Y. H. Tsai, M. H. Yang, and M. J. Black. Video segmentation via object flow. In *CVPR*, pages 3899–3908, June 2016.