Robust Multiple Object Mask Propagation with Efficient Object Tracking

Haibing Ren
YouKu Cognitive and Intelligent Lab
Alibaba Group
haibing.rhb@alibaba-inc.com

Yu Yang
YouKu Cognitive and Intelligent Lab
Alibaba Group
xili.yy@alibaba-inc.com

Xiaoyu Liu
YouKu Cognitive and Intelligent Lab
Alibaba Group
beilin.lxy@alibaba-inc.com

Abstract

In interactive video object segmentation, there are 2 core operations: interactive image object segmentation and video object mask propagation. In our system, the two operations are independent from each other. This paper focuses on the latter one and proposes a method named robust multiple object mask propagation with efficient object tracking. In this method, state-of-the-art tracking technology are utilized to provide region of interest for each target object. It can improve the robustness of mask propagation greatly. Finally, the mask propagation results of all objects are combined together to solve the segmentation overlapping problem.

In our challenge system, we didn't use any pre-computed features. On Davis 2017 validation dataset, our method achieves AUC 0.766 and J&F@60s 0.780. On DAVIS Challenge 2019 interactive video object segmentation, our method took the 4th place via a desktop computer with single NVidia GTX1070. Our final scores are AUC 0.588 and J&F@60s 0.555.

1. Introduction

With the increasing of computation power and development of the algorithms, video object segmentation (VOS) becomes more and more popular in recent years. There are 3 scenarios in Davis Challenge 2019: semi-supervised VOS, interactive VOS, and unsupervised VOS. In industrial area, interactive video object segmentation is especially important for video content generation. The interested objects in the source videos should be segmented to synthesize new videos for advertisement or entertainment. In this case, the object masks with fine boundary are necessary which is impossible for fully-automatic methods. With good interactive VOS technologies and several times of human interaction, we are able to improve the segmentation accuracy step by step.

In the interactive VOS, there are 2 core operations: interactive image object segmentation and video object mask propagation. In our method, the two operations are independent from each other. For the interactive image object segmentation, the user scribbles are transferred into guidance map and input to the interactive CNN along with the RGB images and segmentation result of previous round(optional). For the video object mask propagation, this paper proposes a method named robust multiple object mask propagation with efficient object tracking. It utilized the state-of-the-art tracking technology to improve the mask propagation robustness. In our whole challenge system, we didn't use any pre-computed features. Our method took the 4th place on the interactive track of DAVIS Challenge 2019 via a desktop computer with single NVidia GTX1070. Our final scores on Davis 2017 test-dev dataset are AUC 0.588 and J&F@60s 0.555.

2. Related work

Interactive image object segmentation is to segment the object area in an image with the human interaction. It is an image based object segmentation technology and the goal to get the initial object region for video object mask propagation. The human interaction can be scribbles, bounding box or clicks [1,2]. With such guidance information, the interested object instance can be well segmented.

Video object mask propagation is to obtain the object region given the segmentation results of previous frames. Most mask propagation technologies can be classified into two categories: with or without online learning.

Generally, the online learning based methods fine-tune the model parameters specify for the object instances in the dataset. They use the first-frame annotations and many pre-collected background images to synthesize a fine-tuning dataset with data augmentation methods[1]. These fine-tuning procedures cost very long time. Till the end of last year, nearly all successful methods are with online learning. For example, PReMVOS[4] which is the winner of Davis 2018 challenge on semi-supervised VOS and Youtube-VOS 2018, should fine-tune the networks of object proposal, ReID and mask refinement. The average computation time for each frame is 38s, which means it needs about 1 hour to segment a video clip of 4s. Therefore, they are not suitable for many practical applications, including interactive VOS.

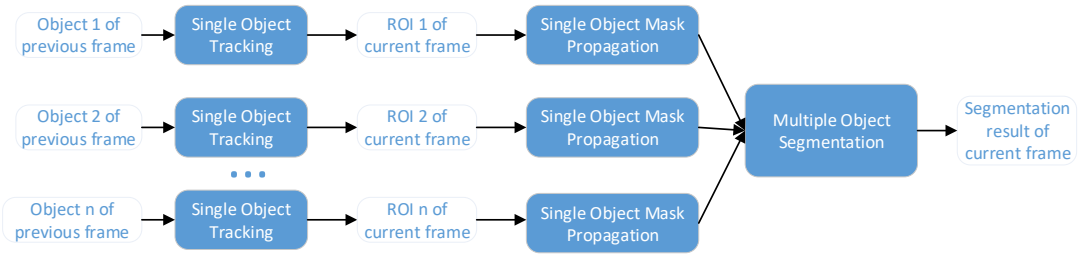


Figure 1. Our mask propagation framework.

From this year, some mask propagation methods without online learning appeared with very good performance. One of the best methods is FEELVOS[5]. For each frame, FEELVOS uses a semantic pixel-wise embedding together with a global and a local matching mechanism to transfer information from the first frame and the previous frame of the video to the current frame. It achieves J&F measure of 69.1% on DAVIS 2017 validation set without any fine-tuning. The computation time is 0.51s/frame. But this method is not robust enough in the interactive VOS. In interactive VOS, the object mask of initial frame is not annotated, but achieved with interactive image object segmentation method. A small error in the object mask of the initial frame will lead to very big mistake after mask propagation of several frames.

In most papers, the two technologies are independent of each other. But in [6], interaction segmentation and mask propagation are conducted by two Convolutional Neural Networks and the two networks are trained jointly to adapt to each other. This method achieved top 1 on interactive t rack of Davis 2018 video object segmentation challenge.

3. Our approach

In our interactive VOS system, there are two different interactive image object segmentation modules: one is for initial frame object segmentation; the other is for later interaction. In both of them, the networks are based on Deeplab v3+[7] and the scribble information is transferred to guidance map. The difference is that the first one has the input of 4 channels (RGB image + guidance map) and the second network's input are of 5 channels (RGB image + guidance map + mask of previous round).

In this paper, we focus on the video object mask propagation algorithm. Its framework is as Figure 1. There are 3 steps in the mask propagation: single object tracking, single object mask propagation and multiple object segmentation.

3.1. Single object tracking

An independent ATOM tracker [8] is applied for each object and the predicted object region in current frame is used to calculate the object region of interest (ROI).

Currently, most state-of-the-art tracking algorithms make multi-scale search and classify the candidates to be the target or not. In the ATOM tracking method, the tracking task is decomposed into two sub-tasks: target estimation and classification. The target estimation component is trained offline and it will predict the Jaccard Index between the target and an estimated bounding box.

Target classification is learned during online tracking and it is to discriminate the target from other objects in the scene by predicting a target confidence score based on backbone features extracted from the current frame.

The backbone networks of both the target classification and estimation tasks are ResNet-18. Therefore, the ATOM tracker can run at 30FPS and the computation time can be neglected in comparison with mask propagation.

3.2. Single object mask propagation

In this module, we designed a fully convolutional network based on Deeplab v3+ with backbone xception_65. The input of first convolution layer is changed from 3 channels to 4 channels, including segmented RGB image of current frame and segmented object mask of previous frame. The output of the network is the segmentation result of current object.

With single object tracking module, a bounding box in current frame will be estimated for each object. With pre-defined enlargement, a ROI of the object is calculated and segmented from the original image. The object mask of previous frame will be segmented with the same method. The RGB region and mask region are resized to the standard size 513*513 and input to the network.

As an accurate ROI is given, the segmentation robustness is greatly improved. The performance of the tracking method is the key for mask segmentation accuracy. In the experiment section, the performance of without tracking module, with ATOM tracker and with DaSiamRPN tracker[9] are compared.

3.3. Multiple object segmentation

In the previous two steps, each object is tracked and segmented independently. Some pixels in the same image will be classified to different objects. In this module, the results of all objects are combined together to solve the object overlapping problem.

Mask propagation module generates a 2D map P_i for

each object *i*. For pixel *x*, the 2D vector $p_i(x)$ is the confidence to be the background and the object *i*. In our method, the 2D vector $p_i(x)$ is normalized as $p'_i(x, k)$:

$$p'_i(x, k) = \frac{e^{p_i(x, k)}}{e^{p_i(x, 0)}} \quad (1)$$

Where *k* (k∈ {0, 1}) is the index of the vector $p_i(x)$. *k*=0 is the background and *k*=1 is the object. After normalization, the background value $p'_i(x, 0)$ is always 1. Then, all the normalized feature map are combined together to segment the final feature map:

$$\begin{cases} \hat{p}(x, i) = \frac{p'_i(x, 1)}{1+\sum_{i=1}^{N} p'_i(x, 1)} \\ \hat{p}(x, 0) = \frac{1}{1+\sum_{i=1}^{N} p'_i(x, 1)} \end{cases} \quad (2)$$

Where *N* is total object number, $\hat{p}(x, i)$ is the probability that pixel x belongs to object *i*. After the feature map normalization, the pixel *x* is classified to object c.

$$c = \underset{0 \le i \le N}{\operatorname{argmax}} \hat{p}(x, i) \quad (3)$$

where c=0 is the background.

4. Training

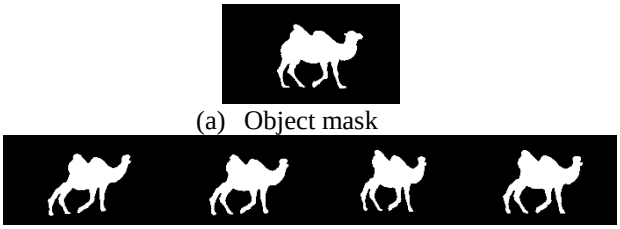
In our mask propagation system, only the network of single object mask segmentation needs to be trained offline. In this network, the input are of 4 channels: segmented RGB image of current frame and segmented object mask of previous frame.

The original Deeplab v3+ network with backbone xception_65 has the input of 3 channels and was trained with MS COCO and PASCAL VOC dataset. For the purpose of our mask propagation, the parameters should be retrained.

In Davis and Youtube-VOS training dataset, there are object ground-truth region of each frame. The object mask of previous frame is always very precise. But in our real application, there are always some errors in segmentation result of previous frame. Therefore, the performance of achieved model will degrade dramatically after several steps of mask propagation. To increase the robustness, we takes two measurements.

- Generate more training samples with data augmentation

  From PASCAL VOC dataset, we can get object instance RGB image and mask ground truth. With affine model, we transfer the object regions to virtual previous masks. The affine model parameters are randomly sampled within some limited range. Then, much more training samples of different object categories can be obtained. An example is shown in the following figure.



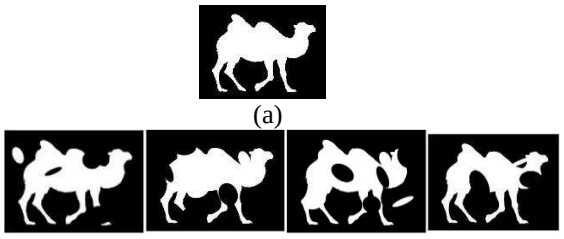
(a) Object mask

(b) (c) (d) (e)

Figure 2. Generate virtual previous masks with object region. (a) is the object mask, (b)~(e) are generated virtual previous masks with random affine model paramters

- Add random noise to the mask of previous frame.

  During training, we add some random noise to the mask of previous frame. Two kinds of ellipse-shape noises with random axis length and orientation are added to the mask image. One is background noise in order to remove some mask region; the other is foreground noise in order to add some non-object region. Some of the examples are as the following figure. The background noises are in black color and foreground noises are in white color.



(a)

(b) (c) (d) (e)

Figure 3. Add noise to previous mask during training. (a) is the ground-truth of object mask. (b)-(e) are generated masks with different noises. In training procedure, (b)~(e) are used for training.

5. Experiments

Davis, Youtube-VOS and PASCAL VOC dataset are used for training the model parameters. On Davis 2017 validation dataset, our method achieves AUC 0.766 and J&F@60s 0.780. In Davis Challenge 2019 interactive video object segmentation, our final scores are AUC 0.588 and J&F@60s 0.555. Some of the segmentation results are as Figure 4.

Figure 4. Some segmentation result on Davis 2017 validation dataset.

Some ablation experiments are carried out to test the performance of the methods with/without object tracking and different tracking methods. The result is as following table:

Performance/ algorithms	Without Tracking	ATOM Tracker	DaSiamRPN Tracker[9]
AUC	0.635	**0.766**	0.702
J&F@60s	0.670	**0.780**	0.731

Table 1. Ablation experiment result on different tracking strategies. Performance is on Davis 2017 validation dataset.

The results in table 1 shows tracking technologies are very important for segmentation accuracy and better tracking method may achieve better performance.

We also compare our mask propagation method with FEELVOS algorithm in the interactive VOS system. With our interactive image object segmentation modules, the FEELVOS 's accuracy on Davis 2017 validation dataset is AUC 0.521 and J&F@60s 0.547 with 8 times of interaction and mask propagation. The poor performance is caused by un-precise masks of interaction frames. The object masks of interaction frames are achieved via our interactive image object segmentation method. Therefore there exists some errors on the initial object masks. With several propagation steps, the error becomes very large.

6. Conclusion

In interactive video object segmentation, the two core technologies are interactive image object segmentation and video object mask propagation. In this paper, we focus on video object mask propagation and propose a robust multiple object mask propagation method. In our method, there are three modules: single object tracking, single object mask propagation and multiple object segmentation. Our method achieved 4th place in Davis challenge 2019 interactive object segmentation.

With the analysis of our experimental results, a lot of error was caused by wrong tracking results. In next step, we will make a better tracking solution. More over, only the previous frame information is utilized in our mask propagation method. Compared with the methods with world top performance, the information on the interaction frames and more previous frames should be combined. This will be our another direction in the future.

References

[1] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, Thomas Huang. Deep GrabCut for Object Selection. arXiv preprint arXiv:1707.00243, 2017

[2] Zhuwen Li, Qifeng Chen, Vladlen Koltun. Interactive Image Segmentation With Latent Diversity. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 577-585

[3] A. Khoreva, R. Benenson, E. Ilg, T. Brox, and B. Schiele. Lucid data dreaming for object tracking. In arXiv preprint arXiv: 1703.09554, 2017.

[4] J. Luiten, P. Voigtlaender, and B. Leibe. PReMVOS: Proposal-generation, refinement and merging for video object segmentat ion. arXiv preprint arXiv:1807.09190, 2018.

[5] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam. FEELVOS: Fast End-to-End Embedding Learning for Video Object Segmentation. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, Seon Joo Kim. Fast User-Guided Video Object Segmentation by Deep Networks. CVPR 2018 Workshops of DAVIS Challenge on Video Object Segmentation.

[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611, 2018.

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, Michael Felsberg. ATOM: Accurate Tracking by Overlap Maximization. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[9] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, Weiming Hu. Distractor-aware Siamese Networks for Visual Object Tracking. European Conference on Computer Vision (ECCV), 2018