Video Object Segmentation using Visual Saliency and Optical Flow

Toan-Anh Nguyen, Xuan-Son Trinh, and Minh-Triet Tran

Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
{ntanh, txson}@apcs.vn
tmtriet@fit.hcmus.edu.vn

Abstract. Video object segmentation is a problem that has a variety of applications such as automatic video matting and 3D reconstruction. However, there is not much focus on the subject comparing to problems like object recognition and image segmentation. In this thesis, the authors seek to propose a novel approach for doing video object segmentation using salient object detection as the main segmentation method, with mask propagation by optical flow and disappearance and reappearance detection using object detection. The approach takes advantage of state-of-the-art methods and can be easily adapted to newer and better methods. Evaluations performed on the DAVIS 2016 dataset for single-mask-object segmentation show that the proposed approach is comparable to state-of-the-art methods on the same problem.

Keywords: Video Object Segmentation, Visual Saliency, Salient Object Detection, Optical Flow, Object Detection.

1 Introduction

Semantic segmentation is the process of partitioning the input image into different meaningful segments that belong to any of the predetermined classes, turning the image into a representation that is easier to analyze [1]. The core difference between semantic segmentation and traditional image segmentation, i.e. the partitioning of the image into sets of pixels (super-pixels) of similar characteristics, is that the segments must now have meanings. For traditional image segmentation, the output is a set of super-pixels in which individual pixel contains similar intensity, color, or other characteristics. The purpose of this is to tell the location of objects and boundaries, so the segments themselves are merely patches of pixels sharing certain characteristics. With semantic segmentation, on the other hand, what we need to do is dividing the image into regions with specific meanings. Thus, the pixels in each segments now no longer share just their own characteristics.

Object segmentation, a semantic segmentation problem, cares about the partitioning of a single object, or objects, in the scene while background information is generally discarded. With video object segmentation, we are given an object of interest and

our job is to segment the object out of a sequence of images, instead of just a static one.

Saliency of an object is defined as the quality of that object to stand out from its neighbors and surroundings [2]. In other words, it can be understood as the quality to draw attention from observers. For computer systems, visual saliency detection let us move from the brute-force approach to processing images, i.e. scanning the every part of an image for the region or object of interests, to rapid candidate selection method [3] [4]. Saliency object detection, the problem of detecting which objects, instead of just regions, in the scene draw focus, also contribute to many different applications, one of which is image/video segmentation [5].

Semantic segmentation has a variety of applications, from automatic video matting to 3D reconstruction. In addition, along with the growth of information technology, multimedia, especially videos, becomes a gigantic source of data. Thus, the applications of semantic segmentation to videos is getting more and more important. It is necessary for efficient algorithms and methods for video segmentation to be researched and developed. However, even with the increasing need for video object segmentation, there is still a shortage for algorithms and methods to resolve it. According Perazzi et al., there is a performance gap between video segmentation algorithms and similar approaches that deal with image segmentation and object recognition. Therefore, those reasons motivates the authors to create a new method for doing video object segmentation. The initial idea is to use salient object detection as a segmentation method because observations show that visual saliency can be done with incredible speed and great accuracy.

The authors' objectives for this thesis thus become proposing a novel approach for video object segmentation that makes use of salient object detection for segmentation. The approach also uses of optical flow to propagate mask information to restrict regions of interest and object detection to detect object disappearance and reappearance. The proposed approach can take advantage of state-of-the-art methods and is flexible enough to adapt to newer and better methods.

The structure of the thesis is as follows: Section 2 presents publications related to the thesis. Section 3 introduces the proposed model for video object segmentation that uses salient object detection. Next, Section 4 describes the experiments that help us build up the final structure for the proposed method and show the results of those experiments. Finally, Section 5 concludes what we have achieved in this thesis and presents future work.

2 Related Work

The Object Flow is published on CVPR 2016 with some characteristics such as specialized for video object segmentation [6], it uses a mask for the initial frame of the video and propagates the mask info among the frames [6]. Therefore, it can take advantage of the temporal information. However, due to their complex and heavy architecture, the OFL has slow running time.

The second is the One-shot Video Object Segmentation. This method is used for video segmentation using mask [7]. However, unlike OFL, OSVOS processes frames independently so it does not take advantage of temporal information [7]. The method has a high accuracy with a persuasive result on DAVIS 2016 contest [8].

Another segmentation method that is relevant to our thesis is the Pyramid Scene Parsing Network (PSPNet), it is published on CVPR 2017 and is used for image segmentation, it does not need any mask and it can segment with high accuracy [9]. However, it is only well adapted to learnt objects from the training set.

Last is the Deep Hierarchical Saliency Network, a saliency object detection architecture [10]. This is also the architecture we use in our thesis as our main segmentation method. This method requires no mask and can detect object saliency with high accuracy and real-time speed (24 fps on modern GPUs) [10].

3 Video Object Segmentation with Salient Object Detection

Salient object detection deals with highlighting the most prominent object out of the whole scene. As such, the method works on the whole image. Furthermore, no mask information is needed with salient object detection, since what we are doing is finding and highlighting the most prominent objects out of the scene. Therefore, when applied to video object segmentation, it might be the case that the object we are segmenting is not the actual object of interest. Therefore, we propose using the mask to restrict the region of interest before the segmentation using visual saliency is carried out.

However, with only a mask of the first frame of the whole sequence available, there is no way a salient object detection method can follow the object of interest as it moves in the scene. Thus, a tracking method to update the mask is needed, and we propose using optical flow for this task.

There is one problem left with using visual saliency as a video object segmentation method. Since we rely on salient object detection, even if we specify the region of interest, whether or not the object is really in that region is completely ignored. This may cause the method to segment an entirely different object instead whenever we lose tracking. To counter this, we propose adding a disappearance and reappearance detection mechanism to the method.

4 Experiments and Results

4.1 Dataset

We perform our experiments on the DAVIS 2016 dataset. The DAVIS dataset is a recently created (2016) dataset focusing on video object segmentation. The dataset contains high definition video sequences that covers a wide variety of actions for four evenly distributed classes: humans, animals, vehicles, and objects [11].

The DAVIS dataset is comprised of two datasets. The DAVIS 2016 dataset aims to aid study on single-mask-object segmentation while the DAVIS 2017 dataset focuses

on multiple-mask-object segmentation. The 2016 dataset is fitting for our purpose of testing the proposed methods. First, the content is diverse enough to ensure we cover enough cases when doing the first set of experiments. Second, the accompanying metrics proposed with the DAVIS dataset provides a meaningful and well-defined validation metrics to assess our method. Third, because the DAVIS 2017 dataset is not fully available, it is hard to quantitatively evaluate our method. Therefore, the DAVIS 2016 is chosen for our study.

4.2 Experiments

To implement the object tracking and mask propagation for the method, we conduct an experiment using EpicFlow to generate the optical flows between each pair of consecutive frames in a sequence and use these results to update the masks. Using optical flow, we can indeed propagate mask information. There are, however, three main problems with this mask update method:

- Since only the first mask for each sequence is available, there is not enough information to effectively update the masks. In other words, noise and loss because of the lack of new information affect the update process.
- Very fast information loss when the object in the scene deforms.
- Optical flows "stuck" when there are object occlusions or disappearance.

Most of the problems boils down to the fact that the mask update process only has old information from the very first mask to work with. Therefore, we use the segmentation results at each of the frames instead of just the information from the first mask itself to update.

The result of the change is better mask propagation since new information is added to the update process at each frame, mitigating both noise and loss while letting the supposedly current state of the object be acknowledged. However, using the segmentation results to update raises another problem. If the segmentation result for a frame is bad, the updated next mask will be affected, which in turns affects the segmentation for the next frame.

To avoid the vicious circle, we propose to apply a lower limit on the size of the bounding box generated from the mask. The reason is bad segmentation results only make the bounding boxes for the objects smaller, not larger. This is because we us the bounding boxes from the mask to restrict the region of interest for segmentation and the segmentation results can only be as large as the bounding boxes themselves. Based on two heuristics, we propose a way to determine the lower limit:

- The union of the mask and the segmentation result for a frame when used with optical flow provides better tracking results.
- When the object of interest in a scene changes its size, the smallest size of the object is not lower than 20% the size when the object first appears.

The second heuristic comes from a preliminary experiment in which we take randomly 1000 patches from the DAVIS 2016 dataset and do the segmentation. The

segmentation results are then evaluated. Visualization of the results can be seen in Fig. 1, which illustrates that Good results mostly have object over region ratio fall in the range $[0.2, 0.7]$.

With those heuristic, we choose the lower limit for the bounding box size of a sequence to be the size of the first bounding box of the sequence. The final proposed structure, illustrated in Fig. 2 consists of three different components:

- Segmentation module: in charge of segmenting object out of the scene.
- Optical flow module: in charge of propagating information from the mask through the sequence, helps create new masks and bounding boxes.

Object detection module: detects objects in the scene, in charge of tracking object disappearance and reappearance.

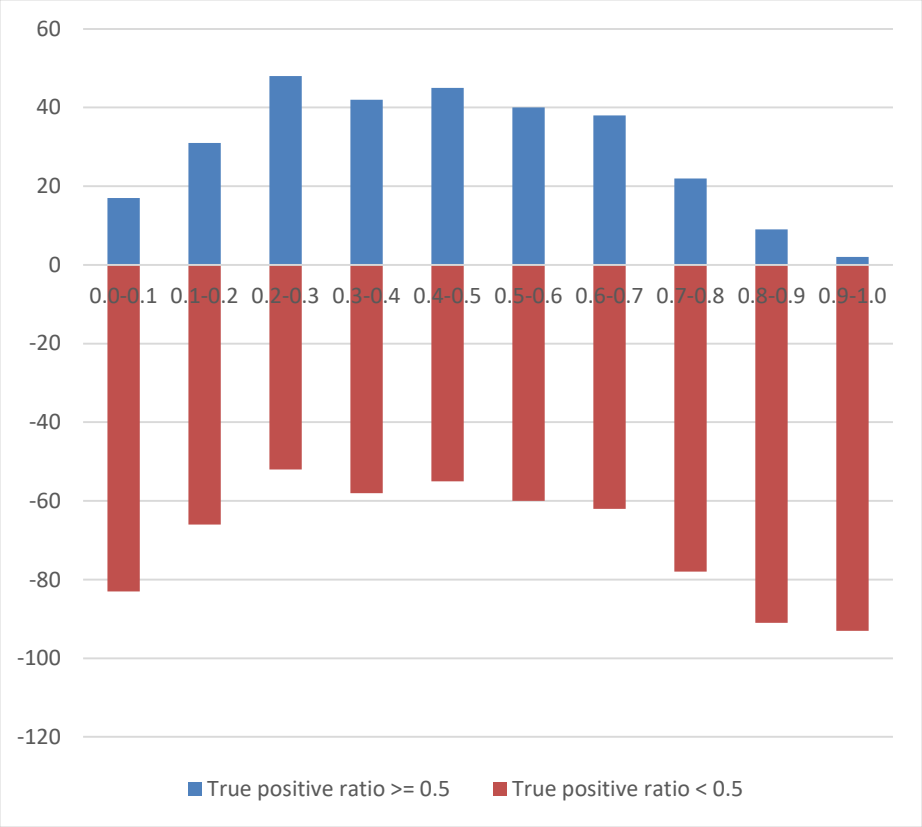

Fig. 1. True positive rate for different object over region ratio

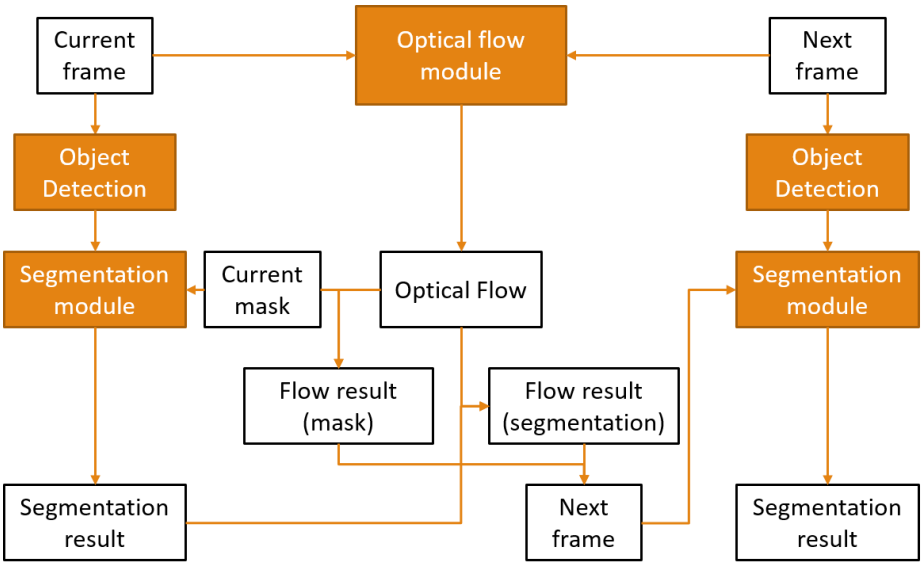

Fig. 2. Final structure for videos salient object segmentation with mask propagation and reappearance detection

4.3 Experimental Results

The experimental results are shown in Table 1 and Table 2, where it can be seen that our final model achieve results comparable to OFL, the fourth ranked method on the DAVIS 2016 *trainval* set as of July 2017. On the *val* set of DAVIS 2016, our method, both with and without object re-identification, is comparable to the third-ranked method VPN while the Ideal Guided DHSNet outperforms the third place.

Table 1. Comparison to other methods on the DAVIS 2016 *trainval* set. Official ranks are as of July 2017. Our results are marked in green (ideal) and blue (current method). Model1 and Model2 respectively represents our method before and after applying disappearance and reappearance.

	J mean	*J* recall	*J* decay	*F* mean	*F* recall	*F* decay	*T*	Official Rank
OSVOS	N/A	N/A	N/A	N/A	N/A	N/A	N/A	-
MSK	**0.803**	**0.935**	0.089	**0.758**	**0.882**	0.095	**0.189**	1
Ideal	0.756	0.902	**0.043**	0.722	0.870	**0.052**	0.348	-
VPN	0.750	0.901	0.093	0.724	0.842	0.136	0.300	2
Model2	0.719	0.842	0.085	0.680	0.800	0.091	0.382	-
OFL	0.711	0.800	0.227	0.679	0.780	0.240	0.224	3
Model1	0.688	0.802	0.112	0.649	0.768	0.119	0.379	-
BVS	0.665	0.764	0.260	0.656	0.774	0.236	0.317	4

Table 2. Comparison to other methods on the DAVIS 2016 *val* set. Official ranks are as of July 2017. Our results are marked in green (ideal) and blue (current method). Model1 and Model2 respectively represents our method before and after applying disappearance and reappearance.

	J mean	*J* recall	*J* decay	*F* mean	*F* recall	*F* decay	*T*	Official Rank
OSVOS	**0.798**	**0.936**	0.149	**0.806**	**0.926**	0.150	0.378	1
MSK	0.797	0.931	0.089	0.754	0.871	0.090	**0.218**	2
Ideal	0.760	0.900	**0.050**	0.722	0.860	**0.050**	0.343	-
VPN	0.719	0.854	0.081	0.678	0.788	0.099	0.376	-
Model2	0.702	0.823	0.124	0.655	0.690	0.144	0.324	3
OFL	0.697	0.825	0.079	0.660	0.765	0.084	0.379	-
Model1	0.680	0.756	0.264	0.634	0.704	0.272	0.222	4
BVS	0.600	0.669	0.289	0.588	0.679	0.213	0.347	5

5 Conclusion and Future Work

5.1 Conclusion

The authors have proposed a method for doing video object segmentation using salient object detection, with optical flow to help in mask propagation and object detection to detect object disappearance and reappearance. We conduct many experiments to study the feasibility of our method and to find ways to improve the current results. The ideal result obtained by incorporating information from the ground-truth encourages us to create a good model. Although it has certain limitations, the proposed method is comparable to state-of-the-art methods.

In conclusion, our thesis provides some insights into applying visual saliency as a semantic segmentation method.

5.2 Future Work

The results for the authors' experiments demonstrate some limitations to the method. There are two main problems with the proposed method:

- For sequences that are subject to occlusion, the method cannot return consistent results.
- Using YOLO for disappearance and reappearance detection is limited in that the object of interest must be consistently labeled as the same class throughout the sequence.

More experiments need to be conducted and more datasets must be used to help us address these problems, as well as improve the current results. Furthermore, experiments and testing are done on the same dataset, testing on different datasets may provide new insights.

References

1. Shapiro, L., Stockman, G.: Computer Vision. Pearson (2001)
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned Salient Region Detection. In : Computer Vision and Pattern Recognition, 2009. CVPR 2009, Miami Beach (2009)
3. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Research 40, 1489–1506 (2000)
4. Navalpakkam, V., Itti, L.: Modeling the influence of task on attention. Vision Research 45, 205–231 (2005)
5. Wang, W., Shen, J., Porikli, F.: Saliency-Aware Geodesic Video Object Segmentation. In : Conference on Computer Vision and Pattern Recognition 2015, CVPR15, Honolulu (2015)
6. Tsai, Y.-H., Yang, M.-H., Black, M.: Video segmentation via object flow. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3899-3908 (2016)
7. Caelles, S., Maninis, K.-K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. (2016)
8. DAVIS: Densely Annotated VIdeo Segmentation. Available at: http://davischallenge.org/index.html
9. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid Scene Parsing Network. In : Conference on Computer Vision and Pattern Recognition 2017, CVPR 2017, Honolulu (2016)
10. Liu, N., Han, J.: DHSNet: Deep hierarchical saliency network for salient object detection. In : Conference on Computer Vision and Pattern Recognition 2016, CVPR 2016, Las Vegas (2016)
11. Perazzi, F., Pont-Tuset, J., McWilliams, B., Gool, L., Gross, M., Sorkine-Hornung, A.: A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In : Conference on Computer Vision and Pattern Recognition 2016, CVPR 2016, Honolulu (2016)