Topic J - STCN Video Segmentations: Final Report

Saifeddine BARKIA
Telecom-Paris
saifeddine.barkia@telecom-paris.fr

Hamza MEDDEB
Telecom-Paris
hamza.meddeb@telecom-paris.fr

Abstract

The goal of this project is to understand how Space-Time Correspondence Networks (STCN) works. Then, we will reproduce the results obtained by the authors on both DAVIS 2016 and DAVIS 2017 Datasets. After that, we will explore the behavior of the algorithm when we use the output of a state-of-the-art segmentation algorithm (Swin-Transformer and Mask-RCNN) and pass it as the first frame of the algorithm. Finally, we will test the generalization ability of the model by testing on another dataset (something to something dataset).

1. Introduction

Video object segmentation is the task of separating the foreground and the background pixels in all frames of a given video. This task is challenging and very important in many domains. In this report, we tackle the video object segmentation issue in a semi-supervised scenario, in which the target object's ground truth mask is supplied in the first frame and the aim is to estimate the object masks in subsequent frames.

This task is very challenging since the appearance of the object can change completely along with the video and also due to occlusions and drifts. Many related works have been proposed to deal with this problem among which we can state: Propagation-based methods, Detection-based methods, and offline learning (which is the category of our algorithm of interest STCN).

2. Difference between STM and STCN

Before exploring the STCN algorithm, we find it essential to first understand the STM algorithm and its architecture.

2.1. STM

In the STM algorithm [1] video frames are sequentially processed starting from the second frame using the ground truth annotation given in the first frame. During the video processing, we consider the past frames with object masks as the memory frames and the current frame without the object mask as the query frame. Those frames are encoded into pairs of key (to encode visual semantics) and value maps (stores detailed information for producing the mask estimation) through deep encoders using RGB input images and masks. In this approach, The affinity matrix is computed for every object in the video separately which is not efficient especially in terms of memory space. Finally, the memory values are transferred to the query frame and the objects are decoded to generate the corresponding mask.

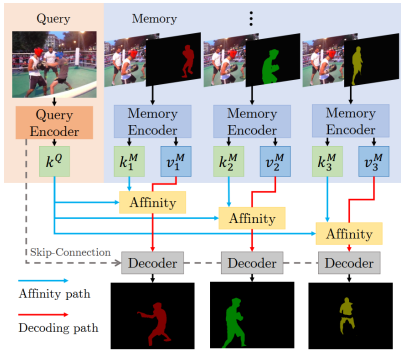

Figure 1. STM architecture

2.2. STCN

For the STCN [2], the authors first encode both the query frame and the memory frame with a key encoder using only RGB inputs then they construct the affinity using these two key tensors. Here, they capture the correspondences with a single affinity matrix without looking at the object masks yet. This is the fundamental difference between STM and STCN. It makes the framework more data robust and computationally efficient. In short, the construction of affinity is redefined to be between frames only so that the key features can be extracted independently without the mask.

Having this affinity matrix, we only now need a simple encoder-decoder structure to transfer mask features. The

final decoding stage generates all the masks.

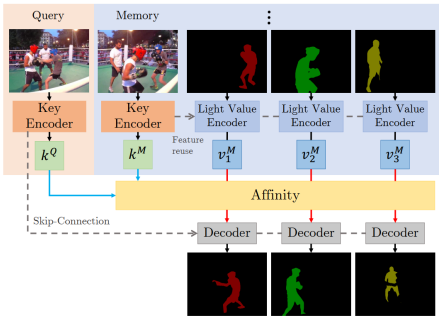

Figure 2. STCN architecture

3. Reproduction of the results

In this section, we are going to report the reported results that we have obtained during our project and compare them to the results reported in the paper

3.1. Results on DAVIS 2016

DAVIS-2016 [3] is one of the most popular benchmark datasets for video object segmentation tasks. In our project, We used the validation set that contains 20 videos annotated with high-quality masks each for a single target object.

Here are some of the reported results of some of the classes in the validation dataset

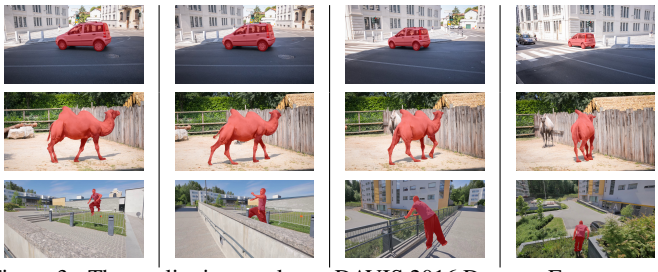

Figure 3. The qualitative results on DAVIS 2016 Dataset. Frames are sampled at important moments. The first row corresponds to the "car-shadow" class. The second row corresponds to the "camel" class. The third row corresponds to the "parkour" class.

We can see visually that from only the first frame annotation, we have been able to track smoothly the object of interest in all the video. We report in Table 1 the quantitative results obtained on all the validation classes.

Our reported results are close to those reported in the paper except for the FPS that's because it is related to the hardware of the computer that we worked on.

Comparison between results		
	Paper results	our reported results
Mean Jaccard index	90.4%	90.3%
FPS	26.9	17.13

Table 1: Quantitative results on Davis 2016

3.2. Results on DAVIS 2017

DAVIS-2017 [4] is a multi-object extension of DAVIS-2016. The validation set consists of 59 objects in 30 videos. We chose to work on this specific dataset to seen if the STCN algorithm is capable of distinguishing the different objects of interest especially when we have occlusions between them.

In Fig 4, we report some of the results that we have obtained in the validation dataset.

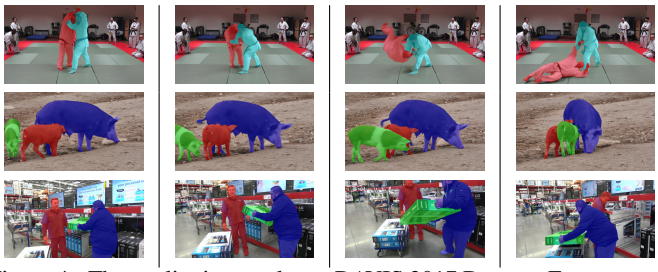

Figure 4. The qualitative results on DAVIS 2017 Dataset. Frames are sampled at important moments (where we have some occlutions between objects). The first row corresponds to "judo" class.The second row corresponds to "pigs" class. The third row corresponds to "parkour" class.

We can see visually that from only the first frame annotation, we were been able to track smoothly all the object of interest in all the videos. Even when we have occlusions in the objects of interest we were able to distinguish between them.We report in Table 3 the quantitative results obtained on all the validation classes for this dataset.

Comparison between results		
	Paper results	our reported results
Mean Jaccard index	82.4%	81.7%
FPS	20.2	10.99

Table 2: Quantitative results on Davis 2017

4. Using state of the art segmentation algorithm to for the first frame segmentation

The ground truth segmentation mask is hard to obtain. It requires a lot of time to annotate manually each pixel. For that reason, we thought of atomizing the task of the segmentation by using state-of-the-art algorithms.

The first segmentation algorithm that we have decided to use is the Swin-transformer [5]. We are not going to explain in detail the architecture of this algorithm since it's not the goal of the project but in short, it is a vision transformer model with a hierarchical way of processing the image. Thanks to that, we can treat every object present in the image regardless of its size in pixels.

We have also tested the Mask-RCNN algorithm [6], which is just an extension of the Faster-RCNN algorithm.

So, instead of passing the ground truth of the first frame to the STCN algorithm, we will provide instead the segmentation generated by the Swin transformer or Mask-RCNN. However, in order to extract only the object of interest(the moving object), we had to apply some post-processing. The main idea was to extract the object with the biggest 8-connected mask. This post-processing is mainly done in two steps. First, we assign a label to each detected mask. Then, we keep the mask with the biggest size as we can see in 5.

Figure 5. The figure on the left is the ground truth, the one in the middle is the objects detected by mask R-CNN, and the one on the right is the post-processed mask.

We report here the segmentation masks obtained by the two state of the art algorithms along with the provided ground truth.

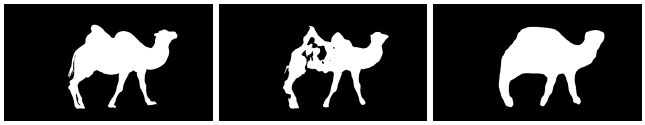

Figure 6. Comparison between the ground truth and the state of the art algorithm for the first frame of "camel". The first figure corresponds to the ground truth. the second corresponds to " Swin-transormer". The third figure corresponds to the " Mask-RCNN"

As we can see in Fig 6, the two masks generated by the state-of-the-art algorithm are not that perfect. This can be explained by the fact that some of the classes of DAVIS dataset don't figure out in the classes of the pre-trained model. However, thanks to the functioning of the STCN algorithm, the masks seem to be better recovered during the video as we can see in 7. In order not to overload the report with comparison examples, we have decided to include the different results in a shared google drive where you can find the different output videos that we have focused on [7].

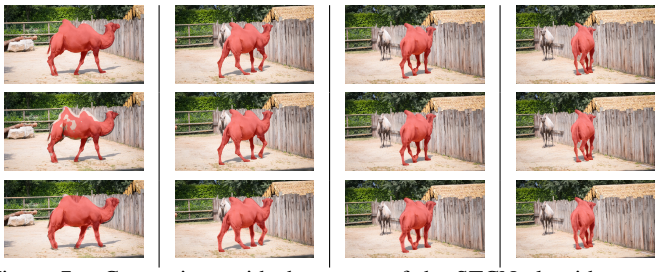

Figure 7. Comparison with the output of the STCN algorithm when fed respectively as the first frame with ground truth, Swin-transformer mask, Mask-RCNN mask.

5. Results on Something to Something Dataset

In section 3, we tested the STCN architecture on a dataset that is constituted with images and their masks. In this section, we will test the model on Something to Something dataset that has frames containing bounding boxes. The coordinates of these bounding boxes can be found in the 'annotations.json' file. Two approaches were tested in this section. The first one is to feed the STCN with the bounding box as the initial mask. The second approach is to use both Mask-RCNN and Swin transformers to generate the initial masks. As we have discussed during the presentation, we had a problem extracting the whole something to something dataset, for that reason, we only worked on a public subset that we have found on GitHub composed of 13 videos [8].

5.1. Using a bounding box as the initial mask

Based on the coordinates provided in the 'annotations.json' file, we generated a rectangle that will be used as the initial mask for the STCN algorithm. Using this approach, we did not expect a good result. Surprisingly, for many videos in the Something to Something dataset a very interesting segmentation was found. For example, figure 8 shows the ability of the STCN to instantly convert a bounding box to a mask.

Figure 8. The left figure is the initial mask. The right figure is the segmentation result of the second frame based on the initialization.

5.2. Using Mask-RCNN and Swin Transformer output as the initial mask

Instead of using a bounding box as the initial mask, we will feed the STCN with the segmentation output of Mask-RCNN or Swin Transfomer. Figure 9, shows an example of an initial mask that was obtained using Swin Transformer.

Figure 9. The figure on the right is the first frame of a video. The one on the right is the segmentation result obtained by Swin Transformer.

In order to obtain a quantitative result, masks need to be converted to bounding boxes. This conversion is done in two steps. First, we extract all the coordinates of the points that compose a mask's boundary. Then, based on these coordinates, we calculate the coordinates of the top left corner of the bounding box which are the minimum of the extracted coordinates on both the x-axis and y-axis, and the coordinates of the bottom right corner which are the maximum of the extracted coordinates on both x-axis and y.

Figure 10. The green line in the middle figure represents the boundary extracted from the mask on the left figure. On the right, is the result of converting the boundary to a bounding box.

We report in Fig 11 the bounding boxes obtained by the STCN algorithm using Mask-RCNN or Swin Transfomer initialization and its corresponding ground truth.

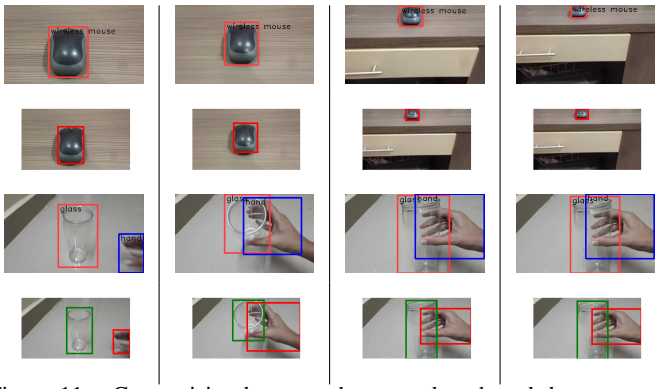

Figure 11. Comparision between the ground truth and the predicted bounding boxes using the detailed approach above for two examples.

	Mean IOU
Mouse	82%
Hand/Glass	75% /85%

Table 3: Quantitative results on the segmented videos of figure 11

The main limitation of STCN is its sensitivity to the initial mask. As we can see in Fig 12, the STCN failed to detect the hand and the marker since it was initialized with only the mask of the watch. One solution that could help to overcome this issue is to feed the STCN with a mask every 10 frames. This way, the STCN will be able to detect objects that are not present in the first frame.

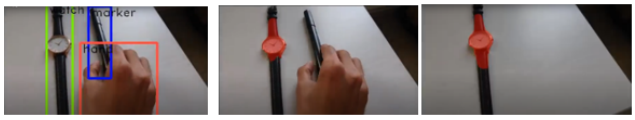

Figure 12. The first figure is the ground truth, the second figure is the mask detected by STCN during the video and the third figure is the initial mask detected by Swin Transformer.

6. Conclusion

Both Mask-RCNN and Swin-Transfomer are computationally expensive. Therefore they cannot be applied to real-time segmentation. Luckily, the STCN solves this issue. . In fact, we only need the segmentation of the first frame to obtain a video segmentation. During this project, we showed that the STCN is robust to the initial mask as we have seen its ability to successfully convert a bounding box to a mask. However, its dependency to the initial mask make it unable to detect objects that appear in the intermediate frames.

References

[1] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9226–9235, 2019. 1

[2] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," *arXiv preprint arXiv:2106.05210*, 2021. 1

[3] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016. 2

[4] 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 preprint arXiv:1704.00675*, 2017. 2

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021. 3

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017. 3

[7] https://drive.google.com/drive/u/0/folders/1Hgyg5CX6tyDG-BvBLI-81VCzXy7N0Cdh. 3

[8] https://github.com/joaanna/something_else. 3