STCN Video Segmentation

LACOMBE Yoach
MVA ENS Paris Saclay

VENARD Paul-Louis
MVA ENS Paris Saclay

Abstract

Spatio-Temporal Correspondence Networks (STCN) [1], presented at NeurIPS2021, are a new method to address the problem of video object segmentation. It refines the concept of STM [6] with simplified encoders and a memory-saving implementation. Despite being a SOTA video segmentation algorithm, STCNs have to be initialized with a first segmented frame. Such a procedure affects the scalability of this algorithm. Thus, in this paper, after reproducing standard STCN results, we tried to automate the preprocessing initialization step using well-known deep segmentation algorithms such as MaskRCNN [2] and PointRend [3], thus replacing a time and energy consuming human segmentation approach. While the qualitative results are promising, we have encountered some difficulties in quantitatively evaluating the resulting videos. Extensive research was done to derive results.

1. Introduction

Video Segmentation is a new field of study in Computer Vision at the hedge of Object Detection and Video Treatment. This task requires a new type of algorithms with light implementation to handle high flow of images and architectures that can process time structured and time consistent data to provides efficient and context-related segmentation. Hence, STM [6] and then STCN [1] provided major advances in that field, with fast inference process and high quality result on different reference datasets. In this perspective, our first work was committed to reproduce the performance results on Youtube VOS [7] 2018 validation set as a confirmation of the promising paper results, and tried to obtain qualitative results on a whole new dataset.

As stated before, the major obstacle to a wider use of such algorithm is the initialization phase. Handly annotated first frames is highly time-consuming and represent an enormous amount of work. The automation of such a task could provide impressive benefits in terms of availability, real-conditions usages and scalability, particularly given the speed and the stellar performances in Object Segmentation of widely spread CNN algorithms, like MaskR-CNN and refined versions like PointRend. Hence in this paper, we developp our attempt to use them in the STCN initialization context, with two datasets, YoutubeVOS and Something-Else [5]. Our finding and results are interesting but they didn't match our expectations. They particularly highlight the need of a common and correlated framework, in terms of classes at least, between Object Segmentation and Video Object Segmentation datasets, as handmade annotations are not limited to MS-COCO class framework : this issue clearly impacted quantitatively our result, even though our automatized algorithm was able to track more objects more precisely. Another issue is the absence of the object of interest in the first frame, skewing the tracking by STCN. We tried to illustrate all these points throughout this paper.

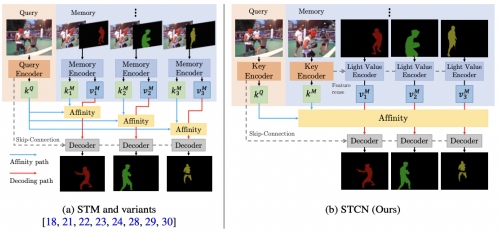

Figure 1. STCN Architecture compared with STM networks

2. Related Works

2.1. STCN

STCN is a refinement of Space-Time Networks (STN). Modeling space-time links, correspondences are established directly between frames, without re-encoding the mask features for every object. Memory coverage is improved thanks to an efficient and single affinity matrix using RGB relations instead of a specific memory bank for every object. Each target object passes through this matrix for feature transfer. With the learned affinity, the algorithm can propagate features from the first frame to the rest of the video sequence, with intermediate features stored as memory.

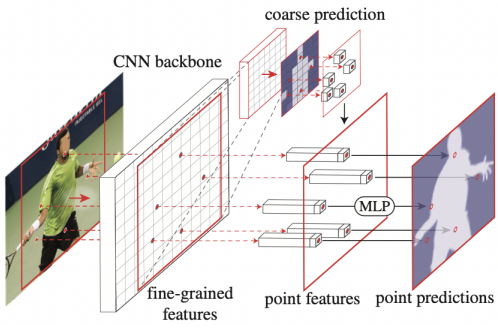

Figure 2. PointRend Design

2.2. MaskRCNN

Mask RCNN are based on Fast RCNN and Convolutional Networks. Faster R-CNN consists of two stages. The first stage, called a Region Proposal Network (RPN), proposes candidate object bounding boxes. It is "backbone" part. The second stage, which is in essence Fast R-CNN, extracts features using RoIPool from each candidate box and performs classification and bounding-box regression. In addition, Mask R-CNN also outputs a binary mask for each RoI. In order to preserve spatial correlation, the ROI proposal backbone was reworked as ROIalign. In this work, we focus on a pretrained network with a Backbone based on ResNet50 and FPN, and a Head with 4 consecutive convolutional networks for classif and segmentation tasks.

2.3. PointRend

PointRend is a new head design add-on. It can be adapted on Mask-RCNN. This new design process segmentation as a rendering problem. Thanks to adaptative subdivision steps, PointRend outputs crisper object boundaries in regions that are over-smoothed by previous methods. To make these predictions, it extracts a point-wise feature representation for the selected points by interpolating the feature map, and uses a small point head subnetwork to predict output labels from the point-wise features.

3. Replication of STCN results

To replicate the STCN paper results, we focused on YoutubeVOS 2018 task, consisting of 474 validation videos, with both seen and unseen categories from the training set. We evaluated the results of STCN on those videos via the evaluation server. We were able to exactly reproduce the results from the papers as seen on figure 8.

4. Automation of the initialization task

4.1. The pipeline

The algorithm pipeline is rather simple, once we got the pre-trained segmentation algorithms MaskRCNN and PointRend, as well as the pretrained STCN. In evaluation mode, STCN takes into input an initial mask and the frames of the video and outputs as many mask images as frames. We thus simply replaced the initial handmade mask by the output of the considered segmentation algorithm taking the first frame of the video.

4.2. Evaluation

4.2.1 Post-processing of the mask

We faced some issues in quantitative evaluation of our results. Indeed, we noticed two main difficulties. First, objects of interest can appear at any frame of the video, whereas we segment on the first frame. Secondly, supervised segmentation algorithms detect only a limited amount of classes, whereas the two datasets considered (YoutubeVOS and Something-Else) are handmade annotated and are tracking all sort of objects. They also don't track every objects of interest of an image, as seen on the 11 in which they only track a few helmets.

Thus, we had to post-process the results, in order to be able to use the Region Similarity and the Contour Accuracy metrics. For a given video with N objects on the initial handmade mask and M objects on the initial automated mask, we tried to associate the i-th handmade object to the closest (in terms of Region Similarity) j-th automated object. If no objects could be associated, the i-th object was not associated at all. This post-processing is clearly not perfect (we still only detect on the first frame, and we are not sure that we associated the right categories together) but it gives a first intuition of the qualitative results.

4.2.2 YoutubeVOS 2018

Qualitatively:

See on figure 5, the initialization masks on an example from YoutubeVOS2018 [7]. Here we notice that the handmade annotations focus on a few helmets, whereas the automated annotations are much more exhaustive and detects frontmost cyclists, as well as their bikes (even though bikes are sometimes partly detected).

See on figure 11, the results of the STCN algorithm on this video with figure 5 masks. The helmets are almost perfectly tracks, only the straps are not tracked. The cyclists tracked thanks to the automated initialization are correctly tracked, but we notice some artefacts around them, maybe due to artefacts from evaluations or due to the performance of STCN, which seems to have difficulties with big masks. STCN seems to have big difficulties with the bikes. We

noticed that except the blue bikes, the algorithm faced some difficulties in tracking the whole bikes, sometimes it even tracked other bikes.

Quantitatively: See figure 9. In that case, Mask-RCNN based algorithm outperformed PointRend based algorithm.

However, no clear conclusion can be taken of these results for the aforementioned difficulties in quantitative evaluation.

Figure 4 gives better insight onto the limitations of this evaluation. Indeed we see that we missed around 0 and 3 objects per videos (comparatively to handmade annotations that tracks objects appearing after the first frame). Moreover we see that the mean region similarity between the N handmade objects and the N automated objects associated with the post-processing is around 0.5, which highlights the second limitation (either the objects were not detected, or only a part of an object was handmade segmented).

The evaluation server also gives per video and per object Region Similarity (J) and Contour Accuracy (F), allowing us to get a better insight into the results. For example (and it is the same with (J) and PointRend) figure 3 which is an histogram of the Region Similarity given by MaskRCNN-based STCN shows us that even though a majority of frame/object associations have around 0 region similarity, the rest is most likely extremely well segmented with a clear bump around 0.9. Our hypothesis is that automated-STCN performs extremely well and that the poor performances on the evaluation server are due to objects appearing after the first frames and object that are only partially handmade segmented.

4.2.3 Something-Else

We applied the algorithm to the Something-Else dataset [5]. It only consists on a small set of 20 object-humans interaction videos with bounding boxes instead of masks.

Qualitatively: When we looked at the results on the 20 videos, the results seems to be fairly efficient. However, we noticed some artefacts of the STCN algorithm, not due to the initalization mask. For example, look at the initial frame on figure 6 and a subsequent frame on 7. We notice that the dial of the watch is also detected between the fingers of the hand, even though the two objects are spatially far away. The only reason we could think of to explain that phenomenon is the dial of the watch is a closed white subset of the image, as well as the space between the finger. Another example is on figure 10, STCN tracked some pixels not related to the pen. The other objetcs were correctly tracked.

Quantitatively: Since we only had bounding boxes as true labels, we produced quantitative results with a modified Region Similarity. Instead of using the classic $\frac{|A \cap B|}{|A \cup B|}$ metric, we used $\frac{|A \cap B|}{|A|}$, meaning it focuses on the boundaries of the mask, i.e this metric approaches 1 when the mask is included in the bounding box. We provided the quantitative results for some videos on table 1. We remark that most of the times, it produces very good results, except for times where no objects are detected in the first frame, or when the objects detected are not among the objects tracked by the bounding boxes.

5. Conclusion

We introduced an unsupervised initialization phase to cover a semi-supervised VOS task. We saw a shift of paradigms in terms of tasks, thus we faced some difficulties in adapting the metrics and judging the efficiency of the new method.

The unsupervised initialization phase is inherently flawed for two main reasons:

1. Objects of interest can appear in any frame at any time. When should we initialize or repeat the initialization phase ?

2. A supervised segmentation algorithm has a limited amount of classes it has been trained on. In the contrary, manually initializing the algorithm allows to concentrate on any kind of objects (see the cyclist helmets).

However,by taking a step back, let's note that STCN fulfils its objectives (supervised VOS) and goes even further because of its efficiency in terms of speed and number of objects it can track. MaskRCNN-STCN and PointRend-STCN could easily be used for human (or every classes the segmentation algorithm has been trained on) detection and tracking. It would be interesting to test this algorithm on these specific purpose, with a dedicated dataset. It would also be interesting to test other segmentation algorithm such as SwinTransformer [4] for Object Recognition.

6. Contributions

Paul-Louis took care of the replication of the STCN results on YoutubeVOS, as well as the inference of the MaskRCNN-STCN and PointRend-STCN on YoutubeVOS. Yoach took care of the post-processing of the masks, the visualization of the masks and graphs, as well as the inference and evaluation of MaskRCNN-STCN and PointRend-STCN on Something Else. The contribution for the slides and the report was equal. Please note that we use a Google Cloud VM to produce our study.

References

[1] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *CoRR*, abs/2106.05210, 2021. 1

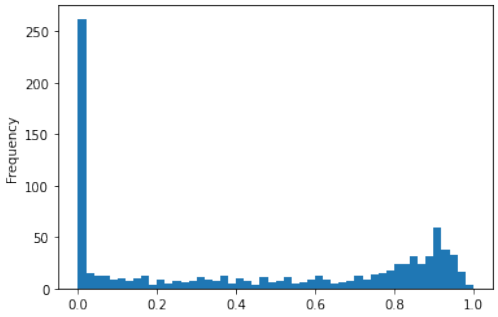

Figure 3. Histogram of Region Similarity per pairs of object/frame with YoutubeVos 2018 validation set, predicted with MaskRCNN-STCN

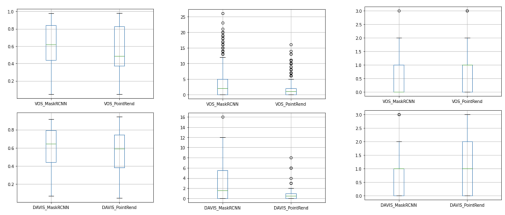

Figure 4. On the left, boxplot of Region Similarity between first handmade masks and first automated masks. On the middle, number of excess objects from automated masks as compared to handmade maks. On the right, number of objects in the handmade masks not detected (that is with a region similarity of exactly 0 - we considered that if region similarity was over 0, the object is somewhat detected.) in the automated masks.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 1

[3] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. *CoRR*, abs/1912.08193, 2019. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *CoRR*, abs/2103.14030, 2021. 3

[5] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. *CoRR*, abs/1912.09930, 2019. 1, 3

[6] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. *CoRR*, abs/1904.00607, 2019. 1

[7] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 1, 2

7. Annexe

Figure 5. Example of the initialization frame (YoutubeVOS2018). From left to right, handmade annotations, MaskRCNN annotations, PointRend Annotations.

Figure 6. Example of the initialization frame (SomethingElse) given by PointRend annotation.

Figure 7. Example of a subsequent frame (SomethingElse) given by PointRend-STCN algorithm.

```
Overall: 0.843389870767
J_seen: 0.831858960005
J_unseen: 0.789511195728
F_seen: 0.879392151872
F_unseen: 0.872797175464
```

Figure 8. Replication of the paper results on YoutubeVOS 2018 after evaluation on the server. The results are exactly the same as the paper.

	Mask R Init	PointRend Init	Manual Init
Overall	0.385833804092	0.313521876924	0.843389870767
J_seen:	0.497220443184	0.400905111934	0.831858960005
J_unseen:	0.245237795564	0.197354650337	0.789511195728
F_seen:	0.530281417835	0.434885223744	0.879392151872
F_unseen:	0.270595559784	0.22094252168	0.872797175464

Figure 9. Results on the YoutubeVOS 2018 validation set, after post-processing of the masks and evaluation on the server. From left to right, MaskRCNN-STCN, PointRend-STCN, handmade-STCN.

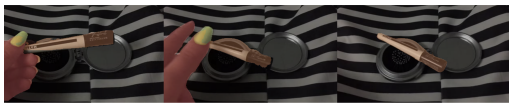

Figure 10. Example of PointRend-STCN results on SomethingElse.

Figure 11. Example of results on YoutubeVOS2018. From up to down, handmade-STCN, MaskRCNN-STCN, Point-RendSTCN.

video_name	mean_jaccard	nb_mask_NotDetected	nb_frames
13201	0.99	0	49
151201	0.05	0	23
2	0.67	5	48
2003	0.41	0	43
22983	0.66	2	52
3201	1.00	0	43
4	0.01	41	42
44862	0.00	45	46
57082	0.00	37	38
6981	0.99	0	23
77005	0.59	0	43
80962	0.00	57	58
862	0.54	14	47

Table 1. SomethingElse with PointRend-STCN.