STCN Video Segmentation

Topic J

Lacombe Yoach, Venard Paul-Louis
MVA ENS Paris Saclay

January 17, 2022

Overview

1.1 Some reminders on STCN

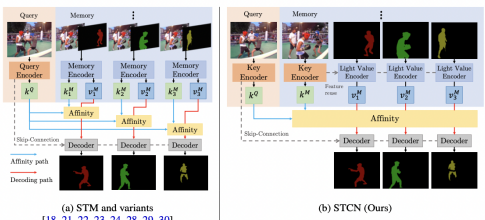

Figure: Scheme of STCN architecture, compared with STM

Space-Time Correspondence Networks (STCN), presented at NeurIPS2021 [Cheng, 2021], are a new method to handle the problem of video object segmentation. It refines the concept of STM [Wug, 2019] with simplified encoders, and memory-saving implementation.

1.2 Challenges

Main Challenge

STCN requires a first annotated (segmented) frame to work

Solution

We could try to automatize this initialization phase, using state of the art Object Segmentation Algorithms:

- MaskRCNN [He, 2018]
- PointRend (refined Network Head) [Kirillov, 2019]
- SwinTransformer

Overview

1. Introduction
2. Replication of Results
3. Automation of the initialization task
4. Conclusion

2.1 Replication of Results

- First, we tried to reproduce the results from the STCN paper.
- We focused on the Youtube VOS 2018 validation dataset.
- We used the official implementation, and infered the results thanks to a Google Cloud VM
- The Datalab test server allowed us to calculate the results
- Metrics based on Region Similarity (J) and Contour Accuracy (F) for both seen and unseen categories.
- Model is pretrained on static Dataset, then on Youtube VOS training and Davis, and finally on synthetic dataset BL30K

2.1 Replication of Results

Method	YouTubeVOS 2018 [78]				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
OSMN [8]	51.2	60.0	60.1	40.6	44.0
RGMP [55]	53.8	59.5	-	45.2	-
RVOS [12]	56.8	63.6	67.2	45.5	51.0
Track-Seg [54]	63.6	67.1	70.2	55.3	61.7
PReMVOS [81]	66.9	71.4	75.9	56.5	63.7
TVOS [82]	67.8	67.1	69.4	63.0	71.6
FRTM-VOS [6]	72.1	72.3	76.2	65.9	74.1
GC [24]	73.2	72.6	68.9	75.6	75.7
SwiftNet* [30]	77.8	77.8	81.8	72.3	79.5
STM [18]	79.4	79.7	84.2	72.8	80.9
AFB-URR [23]	79.6	78.8	83.1	74.1	82.6
GraphMem [16]	80.2	80.7	85.1	74.0	80.9
MiVOS* [21]	80.4	80.0	84.6	74.8	82.4
CFBI [10]	81.4	81.1	85.8	75.3	83.4
KMN [22]	81.4	81.4	85.6	75.3	83.3
RMNet* [29]	81.5	82.1	85.7	75.7	82.4
LWL [7]	81.5	80.4	84.9	76.4	84.4
CFBI+* [83]	82.0	81.2	86.0	76.2	84.6
LCM* [28]	82.0	**82.2**	**86.7**	75.7	83.4
Ours	**83.0**	81.9	86.5	**77.9**	**85.7**
MiVOS* [21] + BL30K	82.6	81.1	85.6	77.7	86.2
Ours + BL30K	**84.3**	**83.2**	**87.9**	**79.0**	**87.3**

Figure: Results on Youtube VOS 2018 Validation Dataset

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

Figure: Reproduction of the results obtained

2.2 Qualitative Testing on a New Dataset

Hand-annotated dataset like YoutubeVOS are unfrequent due to costly annotation work. Something Else is an object recognition dataset composed of object-human interactions. We tried to test the STCN on this new Dataset to obtain qualitative results

Figure: STCN tracking on Something-Else dataset

2.2 Qualitative Testing on a New Dataset

Figure: STCN tracking on Something-Else dataset

Figure: STCN tracking on Something-Else dataset

Overview

1. Introduction
2. Replication of Results
3. Automation of the initialization task
4. Conclusion

3.1 Automation of the Initialization Task

Context

As seen before, STCN is clearly efficient for VOS tasks BUT is highly dependent on

- human intervention
- quality of the initial segmentation

Mask RCNN and more sophisticated models can provide a solution to automatize this initialization process.

3.2 Choices for Automatic Segmentation for First Frame

Mask RCNN

1. Widely used model for Instance Segmentation tasks
2. Backbone : ResNet50 + FPN, ROIAlign
3. Head : Simple 4 consecutive convolutional networks for classif and segmentation tasks

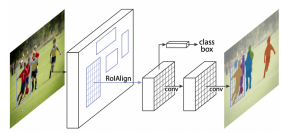

Figure: Mask RCNN Architecture

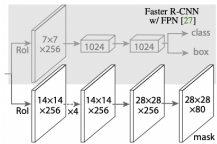

Figure: Mask RCNN head

3.2 Choices for Automatic Segmentation for First Frame

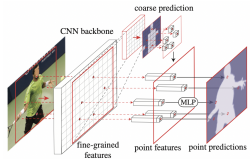

Figure: PointRend Architecture

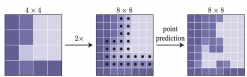

Figure: Upsampling step

PointRend

1. New Head Design Add-on
2. Segmentation as a rendering problem
3. PointRend outputs crisp object boundaries in regions that are over-smoothed by previous methods
4. Adaptive subdivision step

3.3 Comparison between Initial Segmentation Frames

Figure: Initial Segmentation Frames (Manual/MaskR/PointRend)

3.3 Comparison between Initial Segmentation Frames

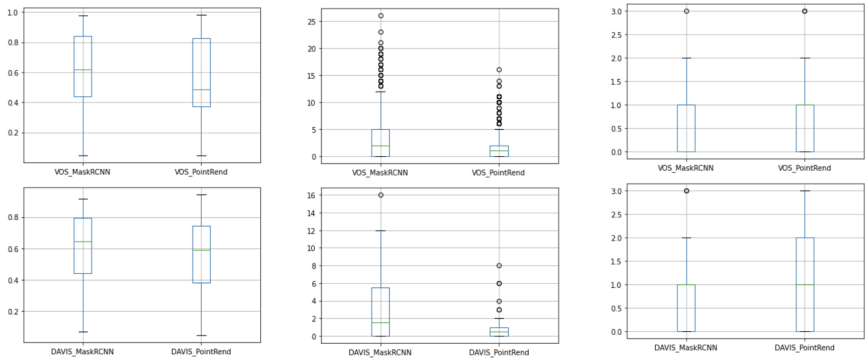

Figure: Box plots of indicators

3.4 Challenges on metrics and evaluation criteria

Many challenges:

1. Handmade initial segmentation only tracks some objects of the frames, sometimes even only parts of these objects. Automated segmentation identifies every instance of object of interest.
2. Object segmentation algorithm are trained only to predict a few classes. How does it influence the performance of STCN ?
3. How to compare tracking ? The problem of YoutubeVOS testing

3.5 Results

Figure: VOS Comparisons (Manual/MaskR/PointRend)

3.5 Results

	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: VOS Comparisons (Manual/MaskR/PointRend)

3.6 Discussion

The task

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.

Initialization Phase

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).

Overview

1. Introduction
2. Replication of Results
3. Automation of the initialization task
4. Conclusion

4. Conclusion

- Pave the way for the better and faster VOS algorithm
- Highlight the need for semi supervisized/unsupervized technics and Datasets
- STCN stays a quick and efficient way of tracking segmentation, even with many objects.

References


Cheng and Tai (2021)

Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation

CoRR abs/2106.05210


Wug and Lee (2019)

Video Object Segmentation Using Space-Time Memory Networks

ICCV


He and Gkioxari (2018)

Mask R-CNN


Kirillov and Wu (2019)

PointRend: Image Segmentation as Rendering

The End