MeNToS : Tracklets Association with a Space-Time Memory Network

Mehdi Miah, Guillaume-Alexandre Bilodeau and Nicolas Saunier

Polytechnique Montréal, Canada

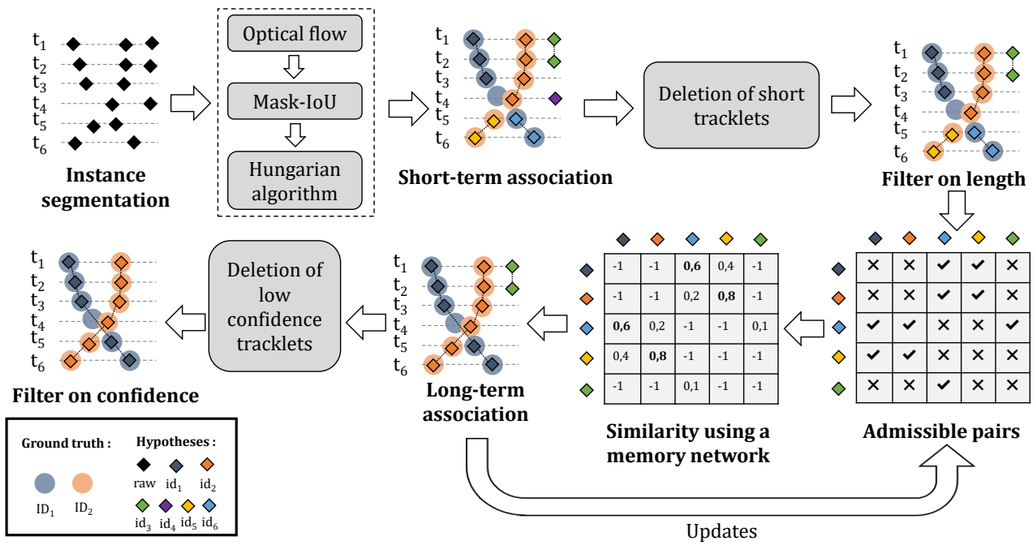
Presentation

RobMOTS is a new tracking challenge evaluated on eight diverse benchmarks. Trackers have to detect all objects from 80 COCO classes without any per benchmark parameters or hyperparameters tuning. Our method is based on **three main steps**: 1) **detections** of all objects of interest as masks, 2) a short-term data association of segmentation masks in consecutive frames to form tracklets and 3) a **greedy long-term data association** of tracklets using a memory network. Our method took the 4^{th} place overall in the challenge.

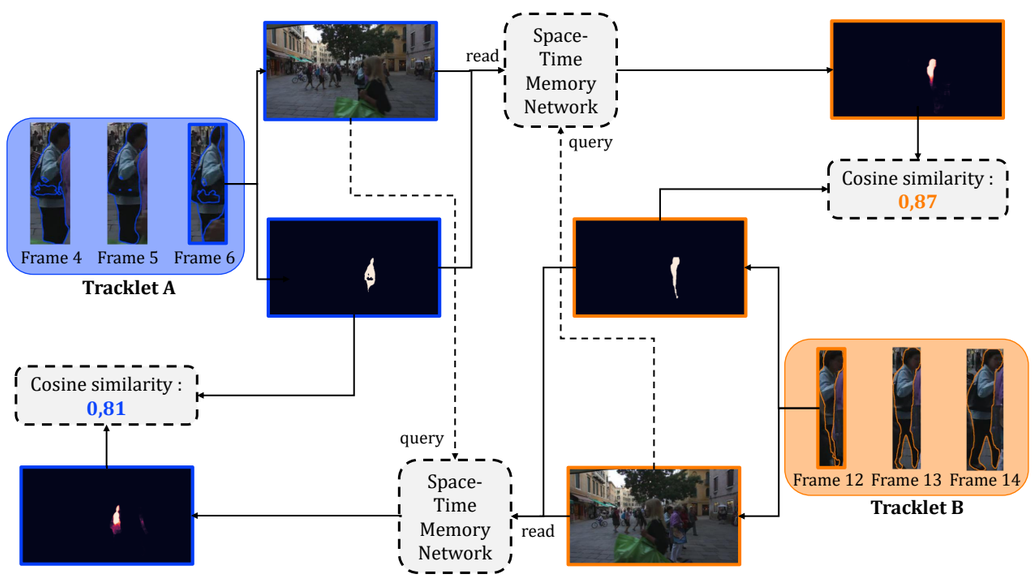
Method

Detections: for all classes, based on the public raw detections with a score higher than 0.50 and bigger than 256 pixels. Masks with a mIoU higher than 0.50 are merged into a multi-class hypothesis.
Short-term data association: detections from consecutive frames are linked to form tracklets, based on the optical flow to predict the future object position, the mask-IoU and the Hungarian algorithm. A non-overlap algorithm is then applied.
Long-term data association: the space-time memory (STM) network [1], a method originally developed for one-shot video object segmentation, is used to propagate some key frames of a tracklet into the past and the future. Depending on the spatial overlap with a mask from another tracklet, both tracklets are merged. The STM network produces a heatmap indicating the probability of the presence of the mask of a tracklet at the pixel level. A cosine similarity is then computed between the heatmap and the actual mask from another tracklet. Finally, tracklets are merged in a greedy manner such that they do not temporally overlap. This constraint on temporal overlap prevents from merging tracklets that appear simultaneously over more than two frames.

General pipeline



Long-term data association : similarity between tracklets



Results

HOTA Scores

Rank	Name	Overall	KITTI	BDD	DAVIS	YT-VIS	TAO	MOTSCha.	OVIS	Waymo
1st	RobTrack	**61.20%**	71.64%	**57.86%**	**56.90%**	**68.32%**	**54.99%**	61.04%	**61.62%**	**57.21%**
2nd	SBT	58.59%	**74.01%**	53.05%	50.26%	64.41%	51.76%	**64.43%**	55.61%	55.23%
3rd	SIA	56.87%	70.76%	53.42%	47.42%	62.70%	49.60%	62.18%	54.76%	54.09%
4th	MeNToS	55.52%	69.71%	52.33%	49.60%	64.19%	39.23%	60.15%	55.56%	53.42%
Baseline	STP	54.35%	66.35%	49.35%	48.21%	62.27%	43.76%	60.35%	52.79%	51.75%

DetA Scores

Rank	Name	Overall	KITTI	BDD	DAVIS	YT-VIS	TAO	MOTSCha.	OVIS	Waymo
1st	RobTrack	**59.43%**	**75.94%**	**48.70%**	**56.84%**	**61.10%**	**52.24%**	68.86%	**59.39%**	**52.35%**
2nd	SBT	55.92%	75.52%	45.57%	49.71%	56.38%	48.54%	**68.91%**	52.86%	49.84%
3rd	SIA	55.83%	75.48%	44.81%	49.75%	55.79%	49.00%	68.89%	52.89%	50.02%
4th	MeNToS	52.38%	71.67%	42.81%	47.68%	56.86%	39.39%	62.88%	53.62%	44.11%
Baseline	STP	55.78%	75.44%	44.81%	49.77%	55.81%	48.90%	68.90%	52.90%	49.73%

AssA Scores

Rank	Name	Overall	KITTI	BDD	DAVIS	YT-VIS	TAO	MOTSCha.	OVIS	Waymo
1st	RobTrack	**64.76%**	68.50%	**70.33%**	**58.04%**	**78.16%**	**59.21%**	54.99%	**65.13%**	63.76%
2nd	SBT	63.07%	**73.41%**	63.58%	51.91%	75.42%	56.71%	**61.06%**	60.05%	62.43%
3rd	SIA	59.81%	67.17%	65.59%	46.31%	72.58%	51.75%	57.08%	58.27%	59.70%
4th	MeNToS	60.80%	68.63%	66.73%	52.69%	74.37%	40.86%	58.43%	59.01%	**65.67%**
Baseline	STP	55.04%	59.24%	56.87%	47.81%	71.63%	41.47%	53.95%	54.17%	55.18%

Competitive results were obtained on all benchmarks except TAO due to its very low framerate: 1 fps (limitation with optical flow).
Our long-term data association can be interpreted as a pixel-level spatio-visual alignment rather than a patch-level visual alignment which is common when the data association is computed using a re-identification network. Experiments show that the long-term data association significantly improves the HOTA score over the datasets used in the challenge.

References

[1] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video Object Segmentation Using Space-Time Memory Networks," in *ICCV*, 2019.

Acknowledgements

NSERC CRSNG