Multitask learning combining detection, segmentation, tracking and forecasting

Vivek Roy Luis E. Navarro-Serment Martial Hebert

Detection, Segmentation, Tracking and Forecasting?

Detection or localization is the task of finding the bounding boxes in the image coordinate frame along with the object class.

Instance Segmentation is the task of drawing segmentation masks for every object differentiating between instances of the same object class.

Tracking is the task of maintaining consistent ids of instances across frames.

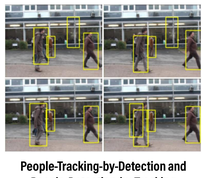
Motion forecasting is the task of predicting the position and size of every object based on its motion history.

- **Detection** and **segmentation** are fundamental tasks for scene understanding
- **Tracking** objects in image sequences helps to understand the dynamics of an object and its interactions with its surroundings.
- **Motion forecasting** is tracking into the future — helps in making the tracker and detector more robust to occlusion.

Motivation

- **Multi-task learning:** Faster R-CNN and Mask RCNN [2] proved that combining the tasks of detection and segmentation can help us perform better in both the task than done individually.
- **Tracking by detection:** Recent surge in the use of detection for tracking. [3,4]
- Tracking and forecasting being related tasks has been done together in the past. [5]

Mask RCNN



People-Tracking-by-Detection and People-Detection-by-Tracking

	Detection	Segmentation	Tracking	Forecasting
Mask RCNN [2]	✓	✓	-	-
People Tracking by Detection [3]	✓	-	✓	-
Track-RNN [4]	✓	-	✓	-
Motion Prediction for People-tracking [5]	-	-	✓	✓
MOTS [1]	✓	✓	✓	-
Ours	✓	✓	✓	✓

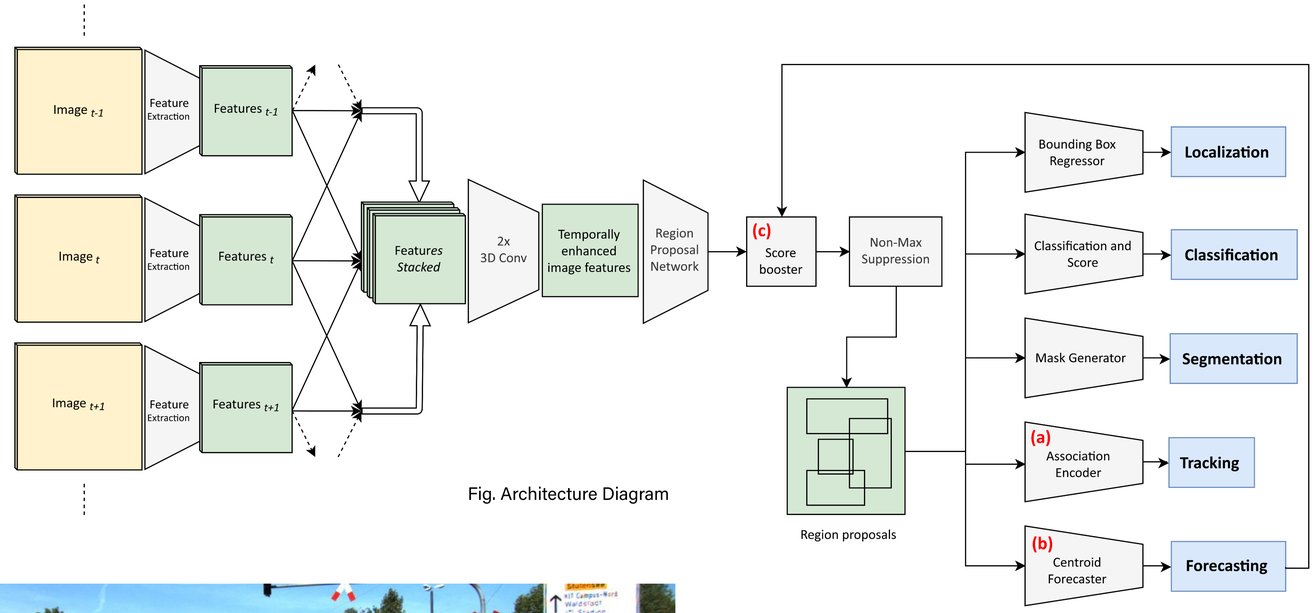


Fig. Architecture Diagram

Results

Fig. showing segmentation masks as well as forecasts. Colored squares represent the location of centroid in the past frames. Solid circles represent the predicted centroid position.

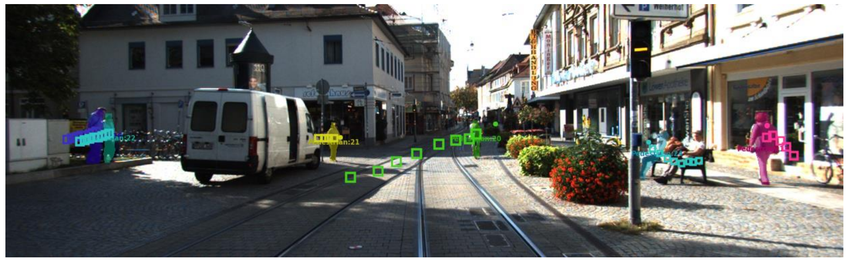


Fig. showing some failure cases. The green bicyclist in the middle does complicated maneuvering relative to the camera motion and the model fails to capture its motion. The man sitting to the right of the person marked in teal on the right half of the image was not detected in the first frame he appeared; since the RPN scores are boosted for the person marked in teal, the man's detection never passes NMS.

		MOTSA [1]	MOTSP [1]	sMOTSA [1]	False Positives	False Negatives	ID Switches
Cars	MOTS	87.8	87.2	76.2	**134**	753	93
	Ours	**91.2**	**88.5**	**78.6**	151	**377**	**46**
Pedestrians	MOTS	65.1	75.7	46.8	**267**	822	78
	Ours	**66.3**	**76.3**	**47.0**	282	**647**	**65**

Approach

Extend the Mask RCNN architecture to incorporate tracking and forecasting.
(Marked with red in the architecture diagram above)

a) **Association module:** An association module added to the Mask RCNN backbone for associating detections across frames.

b) **Forecast module:** A centroid forecasting module added to predict centroid in the next frame.

c) **RPN Score boosting:** The output of forecaster used to boost the RPN scores in the next frame.

Conclusion

- The model reduces the number of false negatives by boosting the RPN scores using the forecast.
- Improved recall helps maintain consistent identities.
- The model captures a good estimate of motion parameters like velocity and acceleration. However, it gives equal importance to the whole past as opposed to giving more importance to the immediate past.

Acknowledgements

I would like to thank the Federation of Indian Chambers of Commerce & Industry (FICCI) for their financial support of this project.

Thank you Dr. Luis Navarro-Serment and Dr. Martial Hebert for your guidance and for this opportunity.

I am also incredibly grateful to Rachel Burcin, Dr. John M. Dolan and Mikayla Trost for their tireless contribution to the Robotics Institute Summer Scholars Program.

Carnegie Mellon University
The Robotics Institute

FICCI

[1] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe, "MOTS: multi-object tracking and segmentation," CoRR, vol. abs/1902.03604, 2019. http://arxiv.org/abs/1902.03604

[2] K. He, G. Gkioxari, P. Dollar, and R. B. Girshick, "Mask R-CNN," CoRR, vol. abs/1703.06870, 2017. http://arxiv.org/abs/1703.06870

[3] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, June 2008, pp. 1–8

[4] K. Fang, "Track-rnn: Joint detection and tracking using recurrent neural networks," 2016.

[5] F. Broz and G. Gordon, "Better motion prediction for people-tracking," 03 2004.