Semantic Segmentation for Pedestrian Detection from Motion in Temporal Domain

Guo Cheng, Jiang Yu Zheng

Department of Computer and Information Science, Indiana University Purdue University Indianapolis, 723 W Michigan St, Indianapolis, IN 46202, USA

IUPUI Center for Visual Info Sensing and Computing

Abstract

Instead of detecting pedestrians from their appearances in 2-D spatial images, which is time-consuming. This paper detects pedestrians along with their motion-directions in a temporal way. By projecting a driving video to a 2-D temporal image called Motion Profile (MP), we can robustly distinguish pedestrian in motion against smooth background motion. To ensure non-redundant data processing of deep network on a compact motion profile further, a novel temporal-shift memory (TSM) model is developed to perform deep learning of sequential input in linear processing time. In experiments containing pedestrian motion from various sensors such as video and LiDAR, we achieve the detecting rate of pedestrians at 90% in near and mid-range on the road. With the super-fast speed and good accuracy, this method is promising for intelligent vehicles.

Keywords – **autonomous driving, motion profile, temporal-to-spatial, semantic segmentation**.

IV. Semantic Segmentation of Pedestrian Motion

For the path planning and autonomous driving on normal roads, we classify pixels on the latest lines in the road profile into three semantic classes: **Pedestrian in motion**, **human standing-still**, **background**.

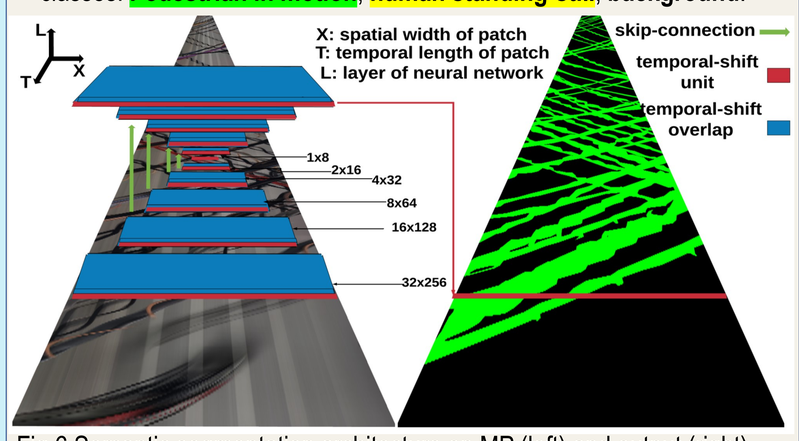

Fig.6 Semantic segmentation architecture on MP (left) and output (right). The network performs semantic segmentation in patch size of 32 × 256 each time, which consists of an encoder-decoder module embedded with skip connections.

V. Sequential Semantic Segmentation with TSM

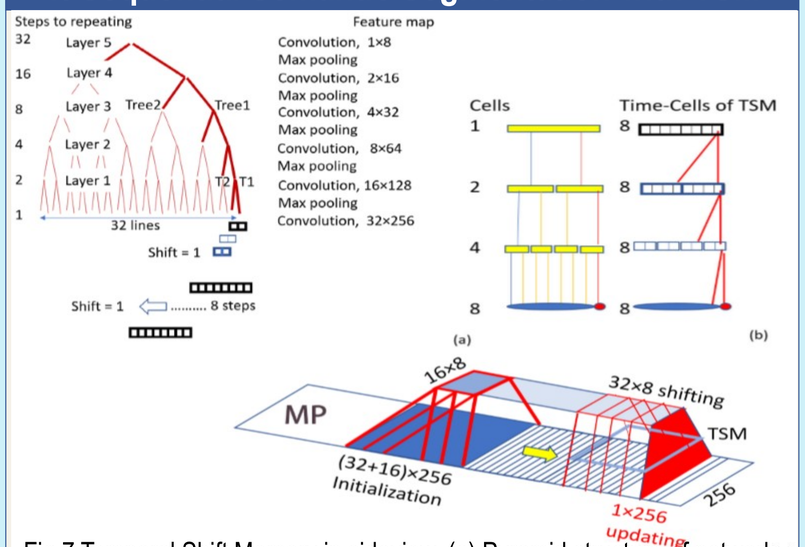

Fig.7 Temporal Shift Memory in sideview. (a) Pyramid structure of network with 2 × 2 pooling and data overlaps. (b) TSM of 3 layers for illustration. Update of cell is done only on the newest nodes (red). (c) TSM avoids repeating calculation on overlapped data for sequential input.

I. Motion Profile for Video Data Reduction

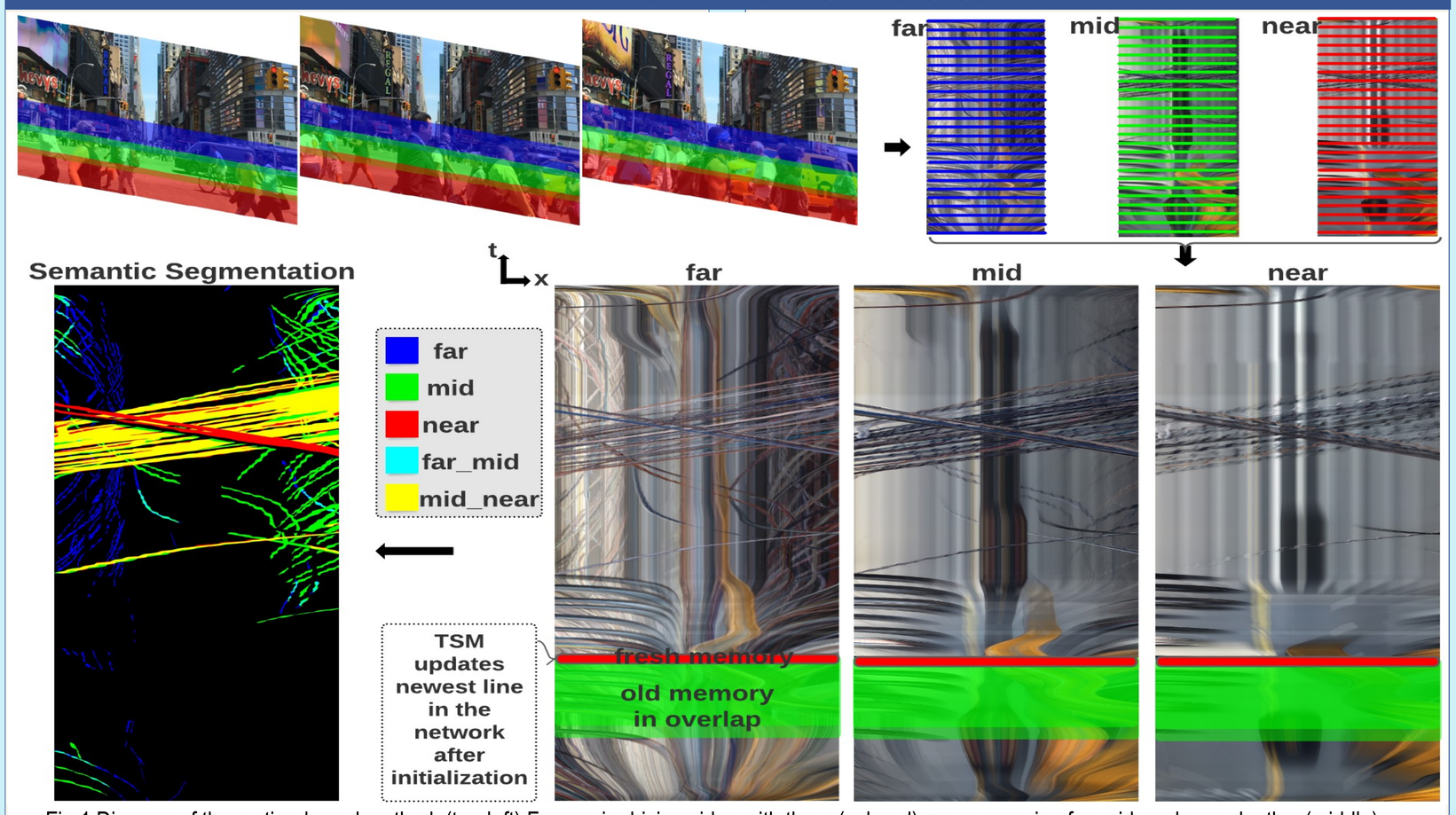

Fig.1 Diagram of the motion-based method. (top-left) Frames in driving video with three (colored) zones covering far, mid, and near depths. (middle) Motion Profiles from three zones in the video formed by averaging pixels vertically in each zone to a line, and then consecutively copying lines into a temporal image. These temporal images contain vehicle, object, and pedestrian trajectories, which will be semantic-segmented through a Temporal-Shift Memory. Results of pedestrian trajectories at far, mid, near depths displayed in three colors (and mixed color in between three depths).

VI. Quantitative Results

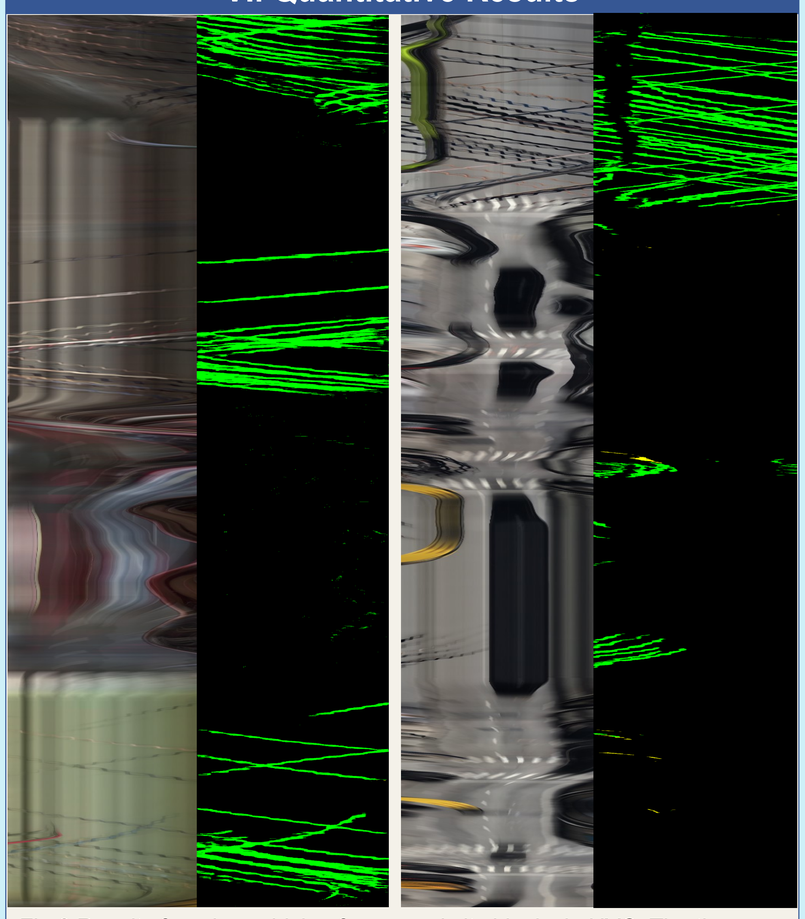

Fig.8 Results from long-driving for several city blocks in NYC. The time aces are upward.

II. Pedestrian Motion in MP

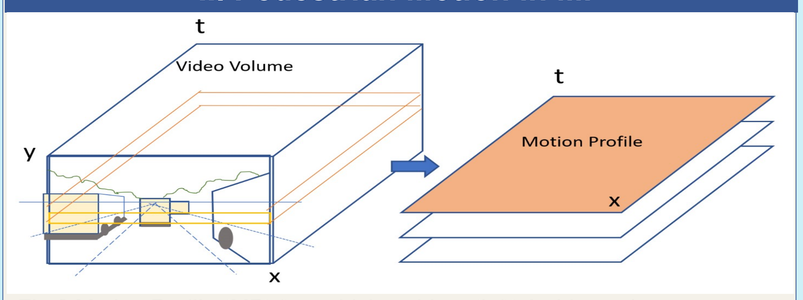

Fig.2 Motion Profile (MP) from driving video. (left) A video volume. (right) A MP image (colored) consists of consecutive pixel lines extracted in video frames within a fixed zone. If the frame of HD video is 720 pixels in height, the data in the MP are 1/720th of video.

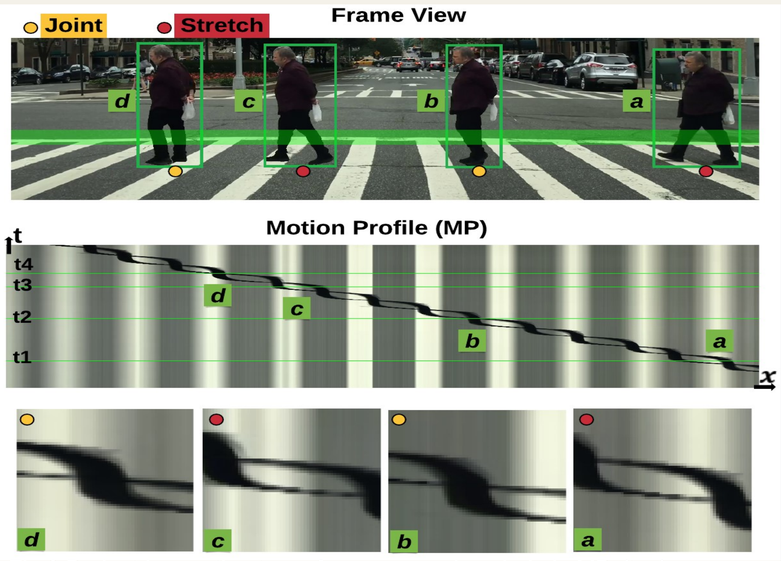

Fig.3 Pedestrian motion trace forms a crossing chain in MP. (top) Pedestrian motion in driving video with a green horizontal belt on legs. (middle) Sequentially projecting the average pixel values in the belt over consecutive frames to generate a temporal image. (bottom) Amplified views of legs at stopping and stepping moments.

III. Patterns of Pedestrian Motion in MP

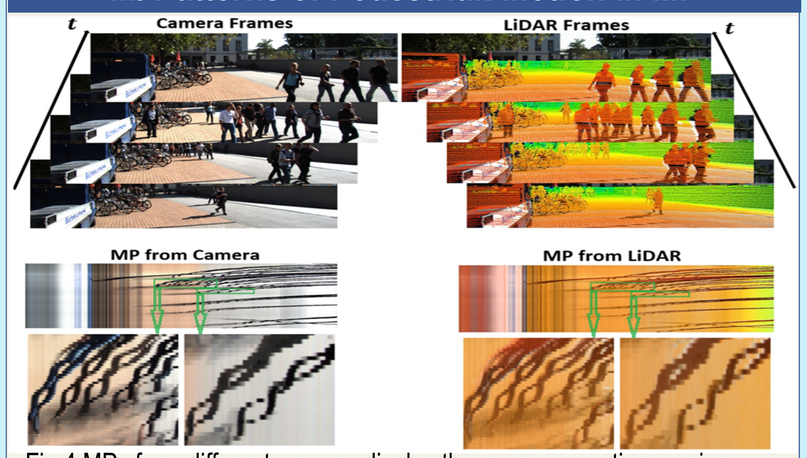

Fig.4 MPs from different sensors display the same semantic meaning.

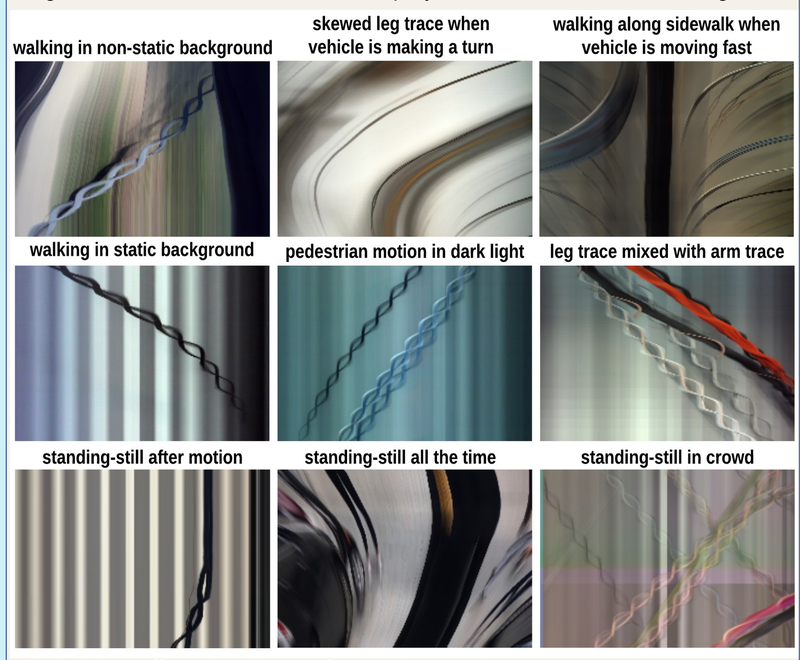

Fig.5 MPs different patterns of pedestrian motion and background.

VII. Evaluation

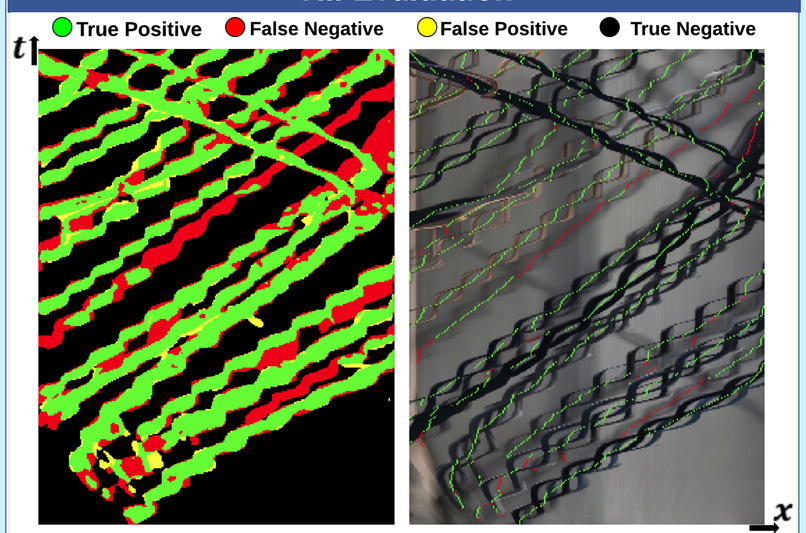

Fig.9 Evaluation of semantic results in a crowd section of MP.

Table I Accuracy of Pedestrian trace at pixel level.

Sensor	Dataset	Precision	Recall	F1	IoU	PA
LiDar	KITTI	0.743	0.650	0.691	0.538	0.964
Camera	Village(TASI)	0.820	0.906	0.861	0.759	0.982
Camera	City(far)	0.658	0.318	0.401	0.269	0.964
Camera	City(mid)	0.794	0.707	0.741	0.600	0.984
Camera	City(near)	0.892	0.928	0.910	0.834	0.996

Table II Accuracy of Pedestrian trace at frame level.

Sensor	Dataset	Precision	Recall	F1	Det.R	PA
LiDar	KITTI	0.929	0.717	0.808	0.685	0.998
Camera	Village(TASI)	0.906	0.957	0.929	0.877	0.999
Camera	City(far)	0.655	0.511	0.553	0.398	0.996
Camera	City(mid)	0.829	0.819	0.819	0.700	0.999
Camera	City(near)	0.946	0.971	0.958	0.920	0.999

Contact the Corresponding Author:

Name: Guo Cheng Email: guocheng@iu.edu