Semantic Segmentation for Pedestrian Detection from Motion in Temporal Domain *Paper ID: 1914

Guo Cheng
Indiana University – Purdue University Indianapolis
Indianapolis IN 46202 USA
guocheng@iu.edu

Jiang Yu Zheng
Indiana University – Purdue University Indianapolis
Indianapolis IN 46202 USA
jzheng@iupui.edu

I. Introduction

In this demo, audience can find a resulting video of pedestrian detection from motion by applying method proposed in the paper. With temporal-to-spatial approach. By condensing driving video into a Motion Profile (MP) as proposed in [1], pedestrians' leg motion forms unique patterns for semantic segmentation as illustrated in the demo video, upon which there are three Motion Profiles extracted at different heights of frame image to capture the motion of pedestrian at far, middle and near distance from the ego-vehicle. Although only 1D data are sampled from each frame, their temporal concatenation to MP shows motion characteristic. Compared to pedestrian detection in video volume, the data sheet of MP is the smallest and the leg motion is less varied than pose, shape, appearance, illumination, etc.

The objective of this work is to find pedestrian motion trajectory at pixel level. By employing a deep learning method, we achieve robust pedestrian detection based on the unique motion patterns in the temporal domain. We conduct semantic segmentation in Motion Profile and improved the accuracy of pedestrian detection by 10%, as compared to existing motion-based works using hand crafted features. This demonstrates that MP is learnable by semantic segmentation and shows the potential degree of using motion alone in pedestrian detection.

To further guarantee the algorithm efficiency of deep learning on the compact MP, another contribution of the paper is to design a temporal-shift memory (TSM) model of the deep network for implementing semantic segmentation without redundant data processing. After computing several initial MP patches, the network only computes the newest line at each layer as MP streamed in. Such a scheme with 1D data from video provides the minimum latency to the road events. By extracting lines from three preset zones in each video frame to cover far, middle and near depth ranges exhaustively, we obtain three MPs in driving video for detecting pedestrians at video rate (30 fps) as well as their trajectories.

In the following section, we introduce related works in pedestrian detection. In Section III, we illustrate the extraction of compact motion profiles from video and various motion patterns of pedestrian walking. In Section IV, an architecture of semantic segmentation network applied to motion profiles is described. In Section V, we focus on the temporal-shift memory for non-redundant calculation of deep network in achieving small latency and dense output. Section VI shows various experiments on different datasets and comparison with other works. Conclusion section summarizes the framework for comprehensive understanding of pedestrian motion.

II. Motion Profile for Video Data Reduction

We sample in each frame at a belt pre-defined below the calibrated horizon to catch the temporal scenes ahead of vehicle. The position of sample belt is set freely to cover a depth range. Multiple belts at different heights with small overlaps can cover close, middle, and far depths, respectively. The belts are set only once after the camera is set on vehicle. To capture the motion trajectory of pedestrians, a belt fixed at leg height in the video catches walking action of human legs in alternative stepping. We condense each belt into 1D array by averaging pixels vertically. Then, the 1D arrays from consecutive frames are copied into a spatial-temporal image, i.e., Motion Profile. The length of MP is the total number of frames in the video, while its width is the same as that of video frame. When a pedestrian is walking, we can observe leg trajectories as a crossing chain. Legs step alternatively in a cycle along the trajectory. Because of the vertical averaging of pixels, we obtained more robust features than pixels, which are strong/long vertical edges in the frame.

Moreover, to avoid collision in driving, we extract three MPs according to far, middle, and close distances. Pedestrians close to vehicle are more dangerous and thus need to be detected accurately and promptly. Pedestrians at far distance have a lower resolution in MP but have chance to be re-identified when ego-vehicle gets closer. Adjacent belts have an overlap of certain pixels, which helps leg detection timely in one of MPs and ensures pedestrian motion observed continuously.

III. Semantic Segmentation of Motion Profile

This work applies deep learning to the motion profile for the first time to enhance this compact data and efficient pedestrian detection. Pedestrian walking shows distinct

patterns in leg alternation against the motions of other objects including background and vehicles as described in [2,3]. Two legs stepping and standing form a chain-type trajectory when a pedestrian is walking.

Moreover, to avoid collision in driving, we extract three MPs according to far, middle, and close distances. Pedestrians close to vehicle are more dangerous and thus need to be detected accurately and promptly. Pedestrians at far distance have a lower resolution in MP but have chance to be re-identified when ego-vehicle gets closer. Adjacent belts have an overlap of certain pixels, which helps leg detection timely in one of MPs and ensures pedestrian motion observed continuously.

IV. Summary of Result

The resulting video is based on driving scene of New York City, it shows pedestrian detection, their motion, and leg span even carrying some bags, wearing skirt, pushing suitcase, in crowds, etc. The bounding boxes upon pedestrians' legs are the projection of instance accuracy from semantic segmentation of pedestrian in Motion Profiles. Upon the frame image, I also displayed the real-time resulting semantic segmentation from motion profiles, which resulting in the bounding boxes. By projecting the semantic segmentation result from Motion Profile into the normal video frame, we can see three boxes detecting pedestrian at far, middle and near. Boxes changing with the leg spans approximately. Multiple boxes are combined to wider ones as crowd. In the resulting video, different colors indicate meanings as following:

Red: pedestrian detection from motion at near distance to ego-vehicle.

Green: pedestrian detection from motion at middle distance to ego-vehicle.

Blue: pedestrian detection from motion at far distance to ego-vehicle.

Compared to other state-of-the-art methods [4-7], our temporal-to-spatial approach presents advantages mainly on: (i) less data; (ii) fast computing time (2ms) in frame advancing, much shorter than YOLO3 (370ms/frame) on the same machine; (iii) preserving a better motion continuity, while YOLO3 leaves some gaps along walking chains, and (iv) higher precision in body width and leg span than bounding boxes. Our method certainly works for surveillance camera and mobile robots with slower motion than vehicles. In addition to video, it is also applicable to LiDAR and infrared cameras, and in night driving when only legs are lit by vehicle headlights.

REFERENCES

[1] M. Kilicarslan and J. Zheng, Visualizing driving video in temporal profile, IEEE Intell. Vehicles Symp., pp. 1263-1269, June. 2014.

[2] M. Kilicarslan, J. Y. Zheng, A. Algarni, Pedestrian detection from nonsmooth motion, IEEE Intelligent Vehicles, 2015, 487–492.

[3] M. Kilicarslan, J. Y. Zheng, and K. Raptis,J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[4] S. Ren, K. He, R. Girshick, J. Sun, and R.-C. N. N. Faster, towards realtime object detection with region proposal netwoks. Proceeding of the 28th NISP, 91-99, 2015, Montreal, Canada.

[5] J. Redmon, S. K. Divvala, R. B. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection. CVPR, 2016, 779–788.

[6] J. Redmon and A. Farhadi, YOLO9000: Better, Faster, Stronger. CVPR, pp. 6517–6525, 2017. [19] J. Redmon and A. Farhadi, YOLO3: An incremental improvement, arXiv preprint arXiv: 1804.02767, 2018.

[7] J. Redmon and A. Farhadi, YOLO3: An incremental improvement, arXiv preprint arXiv: 1804.02767, 2018.