

3D Scene Understanding at Urban Intersection using Stereo Vision and Digital Map

Prarthana Bhattacharyya, Yanlei Gu, Jiali Bao, Xu Liu and Shunsuke Kamijo

Graduate School of Information Science and Technology

The University of Tokyo

Tokyo, Japan

email: prarthana@kmj.iis.u-tokyo.ac.jp and kamijo@iis.u-tokyo.ac.jp

Abstract—The driving behavior at urban intersections is very complex. It is thus crucial for autonomous vehicles to comprehensively understand challenging urban traffic scenes in order to navigate intersections and prevent accidents. In this paper, we introduce a stereo vision and 3D digital map based approach to spatially and temporally analyze the traffic situation at urban intersections. Stereo vision is used to detect, classify and track obstacles, while a 3D digital map is used to improve ego-localization and provide context in terms of road-layout information. A probabilistic approach that temporally integrates these geometric, semantic, dynamic and contextual cues is presented. We qualitatively and quantitatively evaluate our proposed technique on real traffic data collected at an urban canyon in Tokyo to demonstrate the efficacy of the system in providing comprehensive awareness of the traffic surroundings.

Index Terms—autonomous driving, scene understanding, stereo vision, digital map, environment perception, localization

I. INTRODUCTION

Autonomous driving has gathered attention and tremendous interest in the past few decades, with their implementation in commercial cars seeming imminent. Generally, autonomous driving on highways has been demonstrably successful till now. But urban environments because of their complexity, still pose a challenging problem. Challenges include narrow lanes, sharp turns, congested intersections, obstacles, occlusions, blocked streets, parked vehicles, pedestrians, bicyclists and other moving vehicles. Traffic intersections are particularly crucial in this regard. Intersections are called ‘accident-hot-spots’ since misjudging the speed or intent of the surrounding vehicles can easily lead to disastrous collisions. Thus in order to ensure safe operation an autonomous vehicle should be able to continuously and reliably perceive its environment from its sensory inputs. For this purpose, it is evident that it is not just enough to detect the surrounding obstacles across each time step in isolation. The scene has to also be understood by the driving system in reference to the road-lane structure, ego-position, temporal context, as well as the driving task to be performed.

Contribution: To achieve this aim, a vision and map-based approach is proposed in this paper to comprehensively understand the traffic situation at urban intersections. The surrounding traffic participants are detected, tracked, lane localized, while their spatial orientation and temporal behavior are integrated with the road-lane structure.

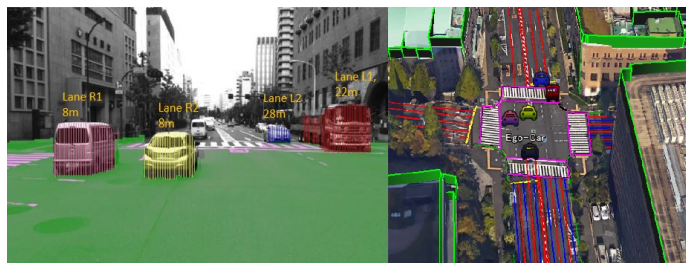


Fig. 1. Understanding a typical urban traffic scene

Fig. 1 shows a typical example of traffic scene understanding at an urban intersection in Tokyo. The figure illustrates stereo-based free-space and vehicle detection and integration of their trajectory with the road structure information included in the 3D digital map of the surrounding. The corresponding flow diagram of the approach is shown in Fig. 2. The stereo camera input images are utilized to generate Semantic Stixels [1], which is basically a way to compactly represent the obstacles present in a 3D scene with rectangles. An initial ego-position estimate is obtained using [2], which is then refined by matching 3D building map data with the 3D stereo input evidence. Map matching also produces a heading direction estimate of the ego-vehicle, which is important to accurately localize surrounding vehicle trajectories. Once the obstacles have been identified and clustered, their dynamics are measured. Finally the Semantic Stixels clusters and their dynamic cues are probabilistically integrated with the map structure and ego-context to provide a 3D understanding of the traffic scene.

II. RELATED WORK

Recently, many research works have utilized vision systems for comprehensive traffic situational awareness. A method proposed in [3] detects and tracks surrounding vehicles while assigning them to their corresponding lanes, and also identifies a leader vehicle which is subsequently used for path planning. However this approach does not make special allowances for complex urban intersections as illustrated in Fig. 1. Another mid-level scene understanding platform is provided by [4], which uses stereo vision and models and tracks obstacles as rectangles with a fixed pixel width. Since they can be clustered together and tracked to impart the notion of objects,

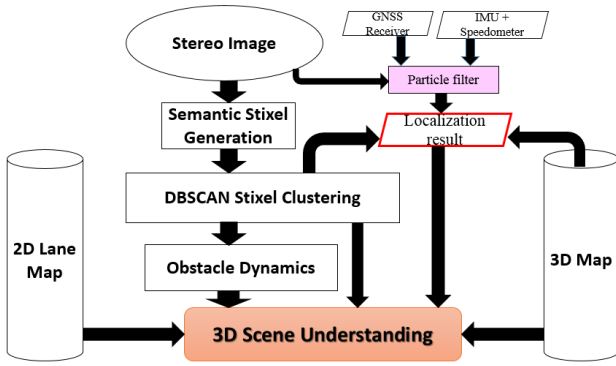


Fig. 2. Flowchart of the Proposed Approach

applications like [5] uses this approach to recognize dangerous situations at roundabouts.

However, it cannot reason about the scene comprehensively owing to a lack of context. An approach in [6] discusses the scene understanding problem from the point of using visual cues like semantic labels, scene flow and occupancy grids to infer the scene geometry and traffic activities, but do not accommodate prior map knowledge into their model. A generative model is proposed in [7] to reason about high-level scene semantics by stressing that there are limited, finite type of traffic patterns in the real-world situations and they can be learned. But the parameters learnt for a particular intersection may not be easily translatable to other intersections. Our approach builds on an extension of [1] and incorporates prior map structure to reason effectively about the 3D scene.

III. FRAMEWORK OF THE SCENE UNDERSTANDING APPROACH

We formulate an approach to probabilistically understand the 3D traffic scene at an intersection. The semantic and geometric cues are obtained from stereo disparity generation algorithms and deep learning based methods. The contextual cues comprise of ego-position and heading direction estimation which are facilitated by the particle filter based integration of various input sensor data. The dynamic cues are obtained from optical flow estimation techniques. These input evidences are discussed in Section (III-A). Section (III-B) provides the framework to fuse these measurements in order to gain a higher-level understanding of the scene in terms of spatial orientation and temporal behavior of the surrounding traffic participants.

A. Input Evidence

1) *Geometric and Semantic Cues of Obstacles*: Firstly the dense disparity images are estimated using DispNet [8]. In the next step, a state-of-the-art and publicly available region-based fully convolutional network (R-FCN) [9] is used to generate pixel level probability scores for different semantic labels. To integrate the geometric disparity cues with semantic labels, a scene model presented in [1] is used to produce a set of Semantic Stixels S_t at time-step t . A single Stixel $s \in S_t$ is defined by a five-dimensional vector $s = [u, v_b, v_t, d, l]$. Here, u is the image column and v_b and v_t mark the base and top

point of the Stixel in image coordinates. The disparity value of the Stixel is d and semantic class is l . Finally, the Semantic Stixels are grouped together in order to join every Stixel with a similar depth and semantic class into the same obstacle. Density-based spatial clustering of applications with noise (DBSCAN) [10] algorithm is chosen by our approach for Stixel clustering since it does not require the predetermination of the number of clusters and can discover clusters with arbitrary shapes. This process assigns a cluster-id $k \in 1, \dots, C$ to each Stixel in the current scene, where C is the number of obstacles at time t . An obstacle $\mathbf{o} \in \mathbf{O}_t$ contains a set of Stixels with the same cluster-id and is defined as $\mathbf{o} = [\{s_i\} : \text{cluster-id}(i) = k]$. The process of integrating semantic and geometric cues from stereo image input data by Semantic Stixels and clustering them to detect obstacles is illustrated in Fig. 3.

2) *Vehicle Self-Localization and Context*: Accurate vehicle self-localization extremely important for scene understanding, and is the key to motion planning and vehicle cooperation. Positioning by Global Navigation Satellite Systems (GNSS) suffer from NLOS propagation and multi-path effects in the urban canyon, while inertial sensors increasingly drift with time. An integrated self-localization system, comprising of GNSS receivers, onboard-cameras and inertial sensors, is proposed in [2] for challenging urban city scenario. This method is modified in this work to include heading-direction correction of the ego-vehicle. In this paper, there are four main sources of positioning, namely Global Navigation Satellite System (GNSS), Inertial Navigation Sensor (INS), stereo-vision and 3D building map. The 3D building map construction has been discussed in [2].

The construction of the 3D map requires the 2-dimensional building footprint, which is provided by Japan Geospatial Information authority, and the Digital Surface Model (DSM) data, acquired from the Aero Asahi Corporation. The height information of the building is included in the DSM data. The 2D map on the other hand, is generated from high resolution aerial images provided by NTT-geospace. Particle filtering is

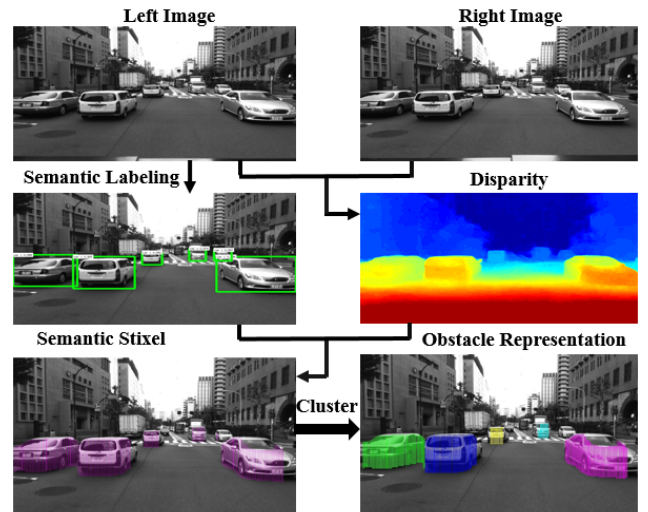


Fig. 3. Integration of Semantic and Geometric Cues for Obstacle Detection

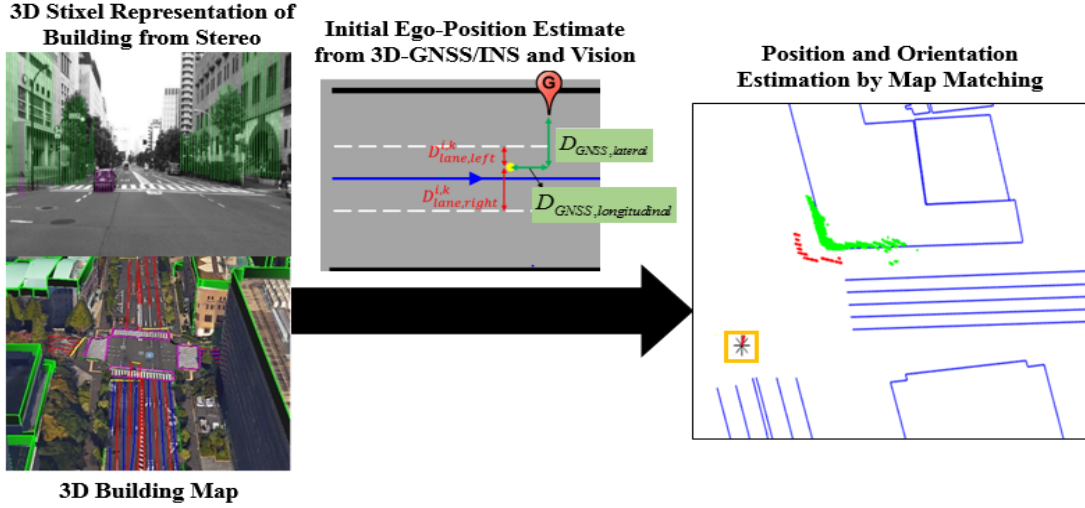


Fig. 4. Estimation of ego-position and orientation. Initial ego-position estimate is found by particle filtering. 3D Building matching calculates orientation.

used to integrate multiple sensor information from GNSS, INSS, vision and 3D map.

INS describes motion of vehicle via the velocity and the heading direction. This information is used for particle propagation in the fusion algorithm. GNSS gives global localization measurement, which can estimate probability of particles. Vision based lane detection perceives the relative distance from the center of vehicle to left white line and right white line. This distance is used to refine the weight of the particles.

In the particle filter, the system state is represented through a set of samples in the inertial frame $\{\mathbf{x}_t = (x_{north}^t, x_{east}^t)\}_{j=1, \dots, n}$. Suppose that a set of n random samples from the posterior probability distribution function $p(\mathbf{G}_{t-1}, \mathbf{V}_{t-1} | \mathbf{x}_t^j)$ is available. GNSS positioning result is \mathbf{G}_{t-1} and \mathbf{V}_{t-1} is the lane detection result at time $t-1$ respectively. Then the weighted average of all particles decides the localization result \mathbf{x}_{t-1} for the time $t-1$. The particle weights are ascertained by equations 1-2 and shown in Fig. 4, where \mathcal{N} represents a normal distribution. σ_{lane}^2 and σ_{GNSS}^2 are empirically chosen.

$$p(\mathbf{V}_t | \mathbf{x}_t^j) = \mathcal{N}(D_{t, left}^j, \sigma_{lane}^2) \cdot \mathcal{N}(D_{t, right}^j, \sigma_{lane}^2) \quad (1)$$

$$p(\mathbf{G}_{lateral} | \mathbf{x}_t^j) = \mathcal{N}(D_{GNSS, lateral}^t, \sigma_{GNSS}^2) \quad (2)$$

$$p(\mathbf{G}_{longitudinal} | \mathbf{x}_t^j) = \mathcal{N}(D_{GNSS, longitudinal}^t, \sigma_{GNSS}^2)$$

The joint posterior probability from which samples are drawn is represented as equation 3 where, is the credibility for GNSS measurement along the lateral direction. With the particle states and positioning result at $t-1$, the particle filter will exclude low-weighted particles, and recursively estimate the localization result \mathbf{x}_t for the ego-vehicle at time t .

$$p(\mathbf{G}_{t-1}, \mathbf{V}_{t-1} | \mathbf{x}_t^j) = \{\gamma \cdot p(\mathbf{G}_{t, lateral} | \mathbf{x}_t^j) + (1 - \gamma) \cdot p(\mathbf{V}_t | \mathbf{x}_t^j)\} \cdot p(\mathbf{G}_{t, longitudinal} | \mathbf{x}_t^j) \quad (3)$$

For estimating the heading direction θ of the ego-vehicle, Normal Distributions Transform (NDT) [11] based map match-

ing is used. Equation 4 represents the spatial matching of the building points from the digital map (shown in green color in Fig. 4) and 3D Stixel representations of buildings (shown in red color in Fig. 4). The idea is to probabilistically align these two spatial vectors in order to recover the parameter θ .

$$\begin{pmatrix} m_{north} \\ m_{east} \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} b_{north} \\ b_{east} \end{pmatrix} \quad (4)$$

The mechanism of position and orientation estimation of the ego-vehicle with respect to the 3D digital map by using GNSS, INS, and stereo-vision information is illustrated in Fig. 4. The black star-shaped symbol denotes the calculated 2D ego-position on the map and a red-dotted line emanating from it represents its heading angle. This provides contextual information for localizing other traffic participants on the map.

3) *Dynamic Cues of Traffic Participants*: Optical flow provides strong cues for temporal scene understanding. In this paper, the popular Lucas-Kanade method [12] is used to calculate the optical flow vectors $\mathbf{U}_{t-1, t}$ for obstacles $\mathbf{o} \in \mathbf{O}_t$, which was obtained in Section (III-A-1).

B. Integration

This section presents a probabilistic fusion technique to temporally combine the semantic, geometric, contextual and dynamic cues and output the digital map position and tracked object-id for all surrounding traffic participants across different time-steps to produce a holistic understanding of the traffic scene.

The position state $\hat{\mathbf{X}}_t$ of obstacle \mathbf{o} with respect to the camera frame of reference is defined as $(\hat{\mathbf{X}}_{north}, \hat{\mathbf{X}}_{east})^T$. In order to obtain an analytic solution to the state, the state transition model is assumed to be linear-Gaussian and is given by equations 5-6. The initial velocity of all objects is set to around 6 m/s empirically. For the consecutive frames, $\mathbf{B}_t = \mathbf{B}_{t-1}$. The measurement model is non-linear and is given by 7. The

process and measurement noise vectors $\boldsymbol{\omega}$ and \boldsymbol{v} are assumed to be Gaussian white noise.

$$\hat{\mathbf{X}}_t = \mathbf{A}_t \hat{\mathbf{X}}_{t-1} + \mathbf{B}_t + \boldsymbol{\omega}_t \quad (5)$$

$$\mathbf{A}_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{B}_t = \begin{bmatrix} \Delta t \cdot v_{\text{north},t} \\ \Delta t \cdot v_{\text{east},t} \end{bmatrix} \quad (6)$$

$$p(\mathbf{Z}_t | \mathbf{X}_t) = p(\mathbf{O}_t, \mathbf{U}_{t-1,t}, \mathbf{x}_t | \mathbf{X}_t) \quad (7)$$

An obstacle $\mathbf{o} \in \mathbf{O}_t$ at time-step t , is assigned the measured image coordinates $(u_{\text{center}}, v_{T,\text{center}})_t$ and disparity value d_{center} from its Stixel cluster. Equation 8 relates the measured obstacle positions at time step t and $t-1$ with the optical flow.

$$(u_{\text{center}}, v_{T,\text{center}})_t = (u_{\text{center}}, v_{T,\text{center}})_{t-1} + \mathbf{U}_{t-1,t} \quad (8)$$

Assuming a pin-hole camera model with baseline b' and focal length f_u [13], the equation for the measurement update at time t is:

$$\begin{bmatrix} u_{\text{center}} \\ d_{\text{center}} \end{bmatrix}_t = \begin{pmatrix} \frac{1}{X'_{\text{north}}} \begin{bmatrix} X'_{\text{east}} \cdot f_u \\ b' \cdot f_u \end{bmatrix} \end{pmatrix}_t + \boldsymbol{v} \quad (9)$$

An Extended Kalman Filter (EKF) [14] is used to obtain $\hat{\mathbf{X}}_t$, the filtered position of $\mathbf{o} \in \mathbf{O}_t$. The final position of the surrounding traffic participants on the digital map is obtained according to the following equation:

$$\hat{\mathbf{X}}_t = \mathbf{R}_y(\theta) \hat{\mathbf{X}}_t + \mathbf{x}_t \quad (10)$$

Here \mathbf{R}_y represents the rotation matrix around y -axis. The temporal association of observation \mathbf{Z}_t to an object \mathbf{o} at time $t-1$ is based on scoring the Euclidean distance to the prediction.

IV. EXPERIMENTAL EVALUATION

The purpose of the experiment is to evaluate our presented approach on real traffic data collected at an urban intersection in Tokyo. In order to evaluate the 3D scene understanding achieved by our proposed technique, we select single-object based tasks as well as a final total scene understanding task that integrates everything. Average accuracy of object detection, semantic labeling, and self-localization are some popular tasks considered for evaluation. We also measure motion states, trajectory, lane information and positioning information of



Fig. 5. Experimental setup to collect real traffic data

TABLE I
QUANTITATIVE EVALUATION

| Object Detection | Detection Rate | False Positive | Frames with False Positive |
|--------------------------------------|----------------|----------------|----------------------------|
| Stixel World [4] + DBSCAN Clustering | 87.2% | 21.6% | 114 |
| Our Method | 92.4% | 0.05% | 27 |

| Object Tracking | MT | ML |
|-----------------|-------|-------|
| Our Method | 42.1% | 10.3% |

| Self-Localization | Lane-Localization Rate of Surrounding Traffic |
|-------------------------------|---|
| Without building matching [2] | 77.3% |
| Our Method | 94% |

surrounding traffic participants from the ego-vehicle. The final task is to integrate both object detection and road map layout to recognize and localize all objects in the road structure across different time frames.

A. Experimental Setup

The experimental setup is shown in Fig. 5. U-blox EVK-M8 GNSS model was used to receive the GPS signals and was mounted on top of the vehicle. Stereo camera is made up of two Point Grey monocular cameras with the baseline of 400mm and set on top of the vehicle facing towards the front. The baseline of the camera is flexible from 300mm to 900mm. For urban city scene, the baseline is set to be 400mm in order to detect both near and far objects. One point grey monocular camera is set inside the vehicle as front view camera to record the ground truth trajectory. CAN data and MEMS-gyroscope data are taken from inertial sensors installed in the vehicle. The stereo data obtained has resolution 15fps, 1024×768 pixel.

B. Experimental Location

The experiment is performed in Hitotsubashi area of Tokyo, where the density and height of buildings is typical for an urban canyon.

C. Quantitative Evaluation

In order to quantitatively evaluate our approach, an 800 frame video at the intersection is chosen, comprising of 80 vehicle sequences. Obstacle detection rate, surrounding vehicle localization rate and tracking metrics Mostly Tracked (MT) and Mostly Lost (ML) [15] are computed as shown in Table I. In the absence of exact ground truth trajectory, we calculate the lane-localization rate of the surrounding obstacles detected in order to estimate the ego-positioning accuracy.

We compare our results to two techniques: results of the DBSCAN clustered Stixel World approach described in [4] and the particle filter based positioning method described in [2] without 3D map matching.

Our results show that object detection based on clustering semantic stixels together with digital map reduces the frames

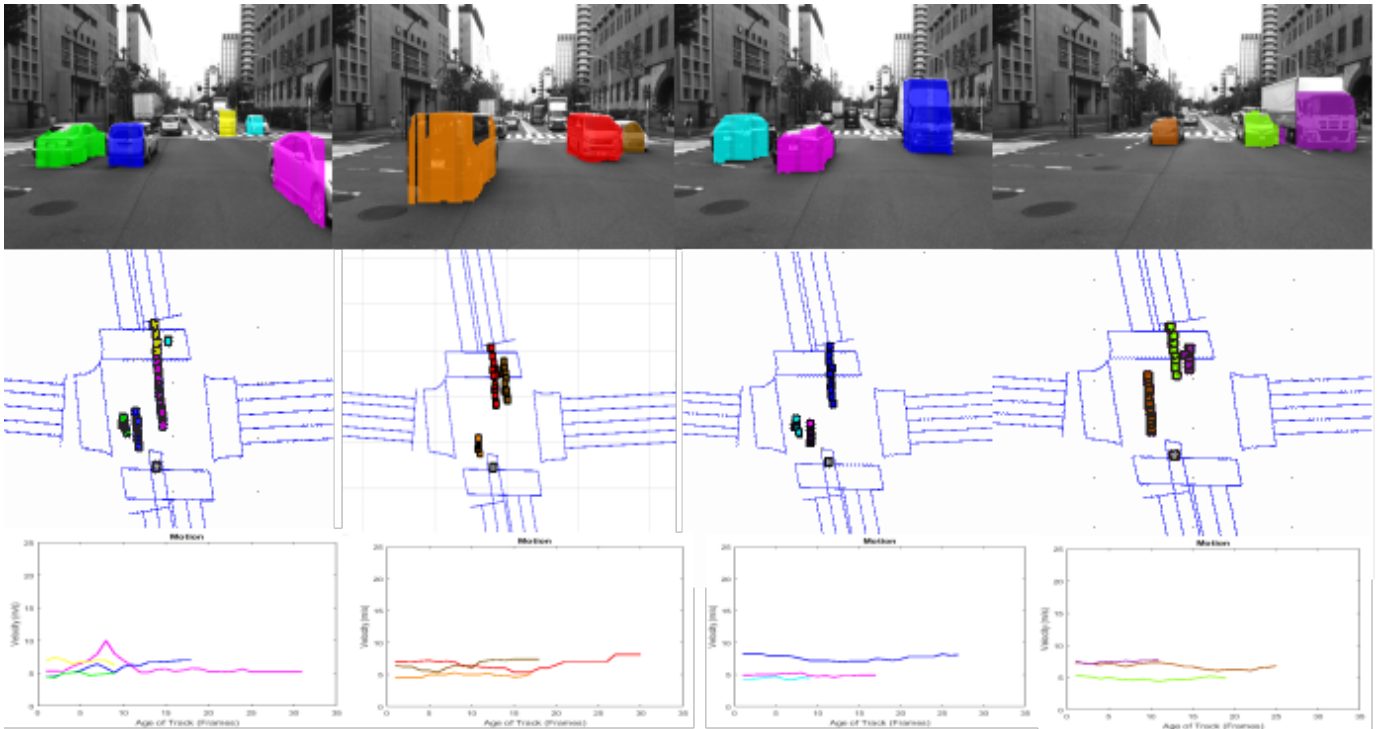


Fig. 6. Qualitative results of the evaluated method. Integration of object segmentation and road map layout for spatial and temporal 3D scene understanding from ego-vehicle platform is demonstrated. The top row shows the obstacle semantic segmentation. The second row localizes the trajectory of the detected obstacles over time on a map. The third row plots their velocities with respect to the age of their corresponding trajectories.

with false positives (FP) drastically. Our comparison also shows that the use of digital map for heading direction estimation helps to reduce the lane-localization error of surrounding vehicles, making the understanding of the system more robust. Tracking results are also presented to describe the temporal correspondence accuracy. We conclude that 3D digital maps help to considerably enhance the scene understanding capabilities of a purely vision based system.

D. Qualitative Evaluation

Fig. 6 illustrates the qualitative results of the proposed method on four example sequences. The top row shows the Semantic Stixel segmentation results of the input images. Accurate and compact representation of the surrounding traffic participants is obtained. The second row shows the integrated spatial and temporal states of the detected obstacles, while their corresponding trajectories are localized on the digital map. The self-localized ego-vehicle is depicted in black. For most sequences the vehicles are assigned to the correct lanes. The vehicle-to-trajectory correspondences are correctly maintained over time. The measured motion state of the surrounding traffic with respect to the age of the trajectory is shown in the third column. This velocity profile is important for the ego-vehicle in order to distinguish between passing and turning cars at the intersection. Overall, an effective perception of the surrounding traffic environment is achieved by this approach.

V. CONCLUSION

In this work, we presented a stereo vision and digital map based framework for robust self-localization and accurate 3D urban traffic scene perception. Additionally a probabilistic fusion of geometric, semantic, dynamic and contextual cues is presented to reason about the scene at a higher level and account for uncertainty in sensor information. Furthermore, precise heading direction estimation of ego-vehicle at turning-intersections and their influence on surrounding traffic localization is addressed. The proposed approach can be used at an urban intersection to answer questions such as: where the ego-vehicle is located on a given digital map; where the surrounding vehicles are located; which car is driving on which street; what their trajectory history is and what the current traffic states are.

Low-cost, close-to-production sensors are leveraged to tackle the challenging accident-prone urban intersection scenario. We quantitatively evaluate surrounding obstacle detection, positioning and temporal correspondence on real urban traffic data and achieve high performance. Qualitative evaluation depicts the integration of the detected obstacles with the road map layout and is shown to provide comprehensive situational awareness. In the future, we plan to use the measured velocity information and lane context of the detected traffic participants to probabilistically infer their intent and predict their maneuver.

REFERENCES

- [1] Lukas Schneider, Marius Cordts, Timo Rehfeld, David Pfeiffer, Markus Enzweiler, Uwe Franke, Marc Pollefeys, and Stefan Roth, "Semantic stixels: Depth is not enough," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 110–117.
- [2] Yanlei Gu, Li-Ta Hsu, and Shunsuke Kamijo, "Passive sensor integration for vehicle self-localization in urban traffic environment," *Sensors (Basel, Switzerland)*, vol. 15, pp. 30199–30220, 12 2015.
- [3] Chunzhao Guo, Kiyosumi Kidono, and Masaru Ogawa, "Vision-based identification and application of the leader vehicle in urban environment," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2015, pp. 968–974.
- [4] David Pfeiffer and Uwe Franke, "Modeling dynamic 3d environments by means of the stixel world," *IEEE Intell. Transport. Syst. Mag.*, vol. 3, pp. 24–36, 09 2011.
- [5] Maximilian Muffert, Timo Milbich, David Pfeiffer, and Uwe Franke, "May i enter the roundabout? a time-to-contact computation based on stereo-vision," in *2012 IEEE Intelligent Vehicles Symposium*, 2012, pp. 565–570.
- [6] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun, "3d traffic scene understanding from movable platforms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 1012–1025, 2014.
- [7] Hongyi Zhang, Andreas Geiger, and Raquel Urtasun, "Understanding high-level semantics by modeling traffic patterns," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 3056–3063.
- [8] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," 06 2016, pp. 4040–4048.
- [9] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.
- [10] Soomok Lee, Daejin Hyeon, Gikwang Park, Il-joo Baek, Seong-Woo Kim, and Seung-Woo Seo, "Directional-dbscan: Parking-slot detection using a clustering method in around-view monitoring system," in *2016 IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 349–354.
- [11] Peter Biber and Wolfgang Straßer, "The normal distributions transform: A new approach to laser scan matching," 11 2003, vol. 3, pp. 2743 – 2748 vol.3.
- [12] Steven Beauchemin and John Barron, "The computation of optical flow.," *ACM Computing Surveys (CSUR)*, vol. 27, pp. 433–466, 09 1995.
- [13] Zhencheng Hu, F. Lamosa, and K. Uchimura, "A complete u-v-disparity study for stereovision based 3d driving environment analysis," in *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, 2005, pp. 204–211.
- [14] Uwe Franke, Clemens Rabe, Hernán Badino, and Stefan Gehrig, "6d-vision: Fusion of stereo and motion for robust environment perception," 08 2005, vol. 3663, pp. 216–223.
- [15] Yu Xiang, Alexandre Alahi, and Silvio Savarese, "Learning to track: Online multi-object tracking by decision making," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.