IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 15, 2018, accepted December 19, 2018, date of publication December 24, 2018,
date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2889320

Robust Vehicle and Surrounding Environment Dynamic Analysis for Assistive Driving Using Visual-Inertial Measurements

YINLONG ZHANG[1,2,3], (Student Member, IEEE), WEI LIANG[1,2], (Senior Member, IEEE), HONGSHENG HE[4], (Member, IEEE), AND JINDONG TAN[5], (Member, IEEE)

[1]Key Laboratory of Networked Control Systems, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China
[2]Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110016, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS 67260, USA
[5]Department of Mechanical, Aerospace and Biomedical Engineering, University of Tennessee, Knoxville, TN 37996, USA

Corresponding authors: Wei Liang (weiliang@sia.cn) and Jindong Tan (tan@utk.edu)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0101200, in part by the International Partnership Program of Chinese Academy of Sciences under Grant 173321KYSB20180020, and in part by the National Natural Science Foundation of China under Grant 61673371, Grant 71661147005, and Grant 61772351.

ABSTRACT Vehicle and surrounding environment dynamic analysis (VSEDA) is an indispensable component of modern assistive drivings. A robust and accurate VSEDA could ensure the driving system reliability in presence of highly dynamic environments. This paper proposes a novel VSEDA framework by fusing the measurements from an inertial sensor and a monocular camera. Compared to traditional visual-inertial-based assistive driving methods, the proposed approach can analyze both the vehicle dynamics and the surrounding environment. Even in the scenario that moving objects occupy a majority area of the scene captured in the image, the proposed method can still robustly analyze the surrounding environment by identifying the static inliers and dynamic inliers, which lie on stationary objects and moving objects, respectively. The theoretical framework consists of three steps. First, the vehicle nonholonomic constraint is applied to pairwise feature matching. For vehicle dynamic analysis, the static inliers are selected by choosing the features with their histogram bins consistent with inertial orientations. Second, for the surrounding environment dynamic analysis, the dynamic inliers are matched through histogram voting, together with the developed part-based vehicle detection model that can segment and match the vehicle regions from the background in image pairs. Finally, both the vehicle dynamics and surrounding environments are analyzed with static and dynamic inliers respectively. The proposed method has been evaluated on the challenging datasets, part of which was collected during rush hours in downtown areas. The experimental results prove the effectiveness and accuracy of the proposed VSEDA.

INDEX TERMS Vehicle and surrounding environment dynamic analysis (VSEDA), assistive driving, monocular camera, inertial measurement unit (IMU), multi-sensor fusion, complex road conditions.

I. INTRODUCTION

Over the past decades, the dramatic improvements on mobile robotics and computer vision techniques greatly enable the popularity of assistive driving, which could liberate the drivers from steering the vehicles in a long term [1]–[3]. The robust and accurate traffic scene analysis is a prerequisite for reliable assistive driving [4], [5]. This is especially important for assistive driving in complex traffic conditions, e.g., during rush hours in downtown areas in extreme weather conditions, so that the autonomous system could yield proper steering commands [6]–[8]. To this end, this paper focuses on the analysis of both vehicle and surrounding environment dynamic analysis.

The images of surrounding environment captured by the vehicle vision system provide abundant surrounding traffic information, such as static inliers (correct feature matches on static scenes) and dynamic inliers (correct feature matches on moving vehicles). Recent works have extensively leveraged various visual feature attributes, (e.g., lane detection [9], feature matching [10], semantic segmentation [11]) to compute

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

vehicle dynamics and simultaneously guide the vehicle on road. It is notable that Scaramuzza [12] has applied the nonholonomic constraint model to estimate ego motion of ground vehicles. The ground-plane assumption dramatically shortens the computational time since only one point is required to fit in the motion model. To take one step further, Sabzevari and Scaramuzza [10] proposed the multi-body motion estimation from the monocular camera. Both the ego motion and surrounding vehicle motions are estimated theoretically in a unified framework.

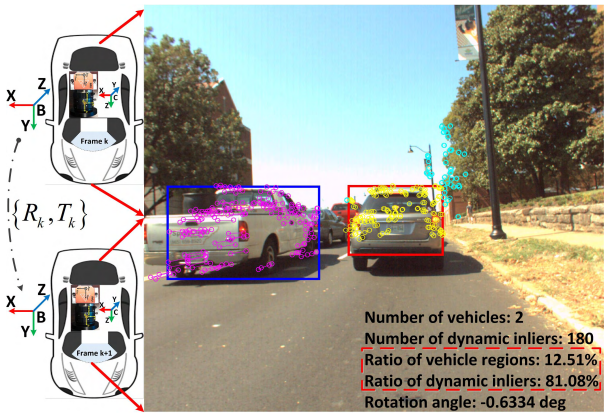


FIGURE 1. Example for visual-inertial based vehicle and surrounding environment dynamic analysis.The image was collected at rush hours in downtown areas of Knoxville. The vehicles are framed by blue and red rectangles. The dynamic inliers in the vehicle regions are marked with pink and yellow circles respectively. The static inliers are marked with cyan circles.

However, the monocular vision system fails to robustly estimate the surrounding vehicle dynamics due to illumination variance, feature matching failures, dynamic motion over-abundance [4], and image blurrings [13]. For example, during the on-road vehicle motion estimation, if the number of dynamic inliers is more than the number of static inliers, such as the case shown in Fig. 1, the pure vision system is unable to recover the vehicle dynamics accurately. The inaccurate or non-robust dynamic estimates on neighbor vehicles, sometimes, will influence the driving maneuvers, which may further incur the unimaginable vehicle accidents. For instance, an unfortunate fatal crash happened to an adult who switched to auto-pilot mode while driving Tesla [14]. The main reason for this accident lies in the vision system failure in detecting the moving tractor-trailer in front during highway crossing. Thus, only the vision system equipped on the vehicle is not sufficient to robustly analyze the vehicle and surrounding environment dynamics.

Complementary to vehicle cameras, inertial measurement units (IMUs) are also able to provide meaningful information for assistive driving. As a proprioceptive sensor, IMU has the heterogeneous and complementary characteristics to cameras, which, by contrast, belong to exteroceptive categories [15]. In general, an IMU commonly consists of tri-axes accelerometers, tri-axes gyroscopes and tri-axes magnetometers. Compared with a visual system, the IMU features in robustness to environmental variations such as illumination changes, image blurs. Also, the IMU is able to capture both severe and subtle movements; so it is suitable to estimate the vehicle states in presence of aggressive or unpredictable motions. Thus, this paper seeks to fuse the visual-inertial sensing modalities to analyze the on-road vehicle and surrounding environment dynamics.

The state-of-the-art VSEDA methods have several potential limitations. Firstly, the vehicle dynamic analysis fails if the major areas of images are taken by moving vehicles. For the example given in Fig. 4 (a) and (b), the estimated motions from camera will be the motions relative to the moving vehicles, rather than to the static scenes; Secondly, the relevant work in analyzing surrounding vehicle dynamics tried to implicitly detect the car regions using heuristic rules, which requires patience and expertise in tuning model parameters [10].

This paper proposes a theoretical framework for analyzing the vehicle dynamics and surrounding environment dynamics using a monocular camera and an IMU. The static inliers are selected through a unified statistical model that considers both inertial cues and visual nonholonomic constraint. Then the dynamic inliers are determined by majority voting upon the vehicle regions localized by the part based vehicle detection model. Ultimately, the vehicle dynamics are derived using the static inliers and the surrounding environments are analyzed using the dynamic inliers. The VSEDA is performed extensively upon the challenging traffic datasets with complex road conditions such as downtown areas during rush hours.

The main contributions of this paper are as follows:

(i) A novel theoretical framework is designed that takes heterogeneous and complementary sensors: IMU and monocular camera to collaboratively analyze both the vehicle dynamics and surrounding environment;

(ii) The inertial guided visual feature labeling mechanism is designed that is able to effectively separate the visual static inliers, dynamic inliers and outliers. The visual static inliers and inertial measurements could be tightly fused in the optimization framework to analyze the vehicle dynamics;

(iii) A vehicle detection model is specifically developed to explicitly detect the moving vehicles in the surrounding environment. The number of vehicles, the ratio of vehicle regions to the whole scene, and the ratio of dynamic inliers are utilized to analyze the surrounding environment.

The rest of the paper is organized as follows. In Section II, the related works on state-of-the-art VSEDA is introduced. The proposed monocular visual-inertial based VSEDA modelling is presented in Section III. In Section IV, the inertial assisted visual feature selection is described in detail. In Section V, the vehicle detection and localization model is given. In Section VI, the vehicle and surrounding environment dynamics analysis are provided. In Section VII, the experimental results and analysis of VSEDA are presented. Finally, the conclusion is given in Section VIII.

II. RELATED WORK

Scholarly works on vehicle dynamics and surrounding environment analysis are extensive. Obviously, to give a comprehensive literature review using a variety of tools (such as Lidar based, radar based, depth camera based, etc) would be unrealistic within the limited pages of this paper. Thus, only the most relevant literatures on visual-inertial based vehicle dynamics analysis and surrounding environment analysis will be reviewed in this section.

A. VISUAL-INERTIAL BASED VEHICLE DYNAMICS ANALYSIS

Basically, state-of-the-arts on visual-inertial based vehicle dynamics analysis can be categorized into two aspects. The first aspect is related to camera viewpoints. From this perspective, they are divided into downward-looking and forward-looking fashions. The second aspect is related to visual error minimization: photometric error minimization or re-projection error minimization.

1) CAMERA VIEWPOINT CONFIGURATION

Vehicle mounted camera is fixed in two ways: camera lens facing downward and camera lens facing forward. In the downward-looking pattern, the ground features are detected and tracked over successive frames. It is assumed that the vehicle moves on the planar ground. Since the height of camera is known a priori, vehicle movement scale ambiguity problem can be significantly resolved [16], [17]. Besides, coplanar ground features are fitted to homography model, which requires fewer samples than that in the epipolar-constraint model. Yet, the major drawback lies in its planar structure constraint, which is not suitable for vehicle pose estimation on uneven terrains, such as bumps and pits. By contrast, in order to capture both the stationary landmarks (e.g., trees, lanes) and dynamic objects (e.g., cars, trucks), the majority of works choose to rigidly fix the camera with its optical axis parallel to the ground. Together with IMU data, vehicle poses can be estimated robustly. The translational component scale is estimated by jointly using inertial linear accelerations and visual landmark positions, either within filtering framework (e.g., MSCKF [18]) or optimization framework (e.g., OKVIS [19])). Besides, the captured images also contain moving objects appeared in the camera field of view, which can also be detected and estimated for further applications [10].

2) VISUAL-INERTIAL MINIMIZATION MODEL

Basically, the visual-inertial based minimization model could be divided into two types: photometric error model and re-projection error model. The photometric type operates directly on image intensities based on dense or semi-dense geometry, such as LSD-SLAM [20], ROVIO [21]. The merit is its full exploitation on all visual information. Hwangbo *et al.* [22] applied the gyroscopic measurements to refine the initialization of the KLT based nonlinear optimization, which increases the possibility of warping parameter convergence. By contrast, re-projection error model estimates 3D geometry from the group of distinctive keypoint feature positions and the inertial measurements. On one hand, IMU helps accelerating and refining the visual feature matching. For example, Kurz and Himane [23] aligned the orientation of local feature descriptors, such as ORB, with the assistance of gravitational components measured from the inertial sensor. On the other hand, the inertial and visual parts could be jointly formulated in a unified objective function within the optimization framework to derive the moving trajectory, such as ORB-SLAM [24] and Manifold-VIO [25]. Comparatively, re-projection error model requires more time to proceed with pose estimations, but achieves relatively more robust results. It is more applicable to large viewpoint pose estimation situations.

B. VISUAL-INERTIAL BASED SURROUNDING ENVIRONMENT ANALYSIS

The surrounding environment analysis denotes the perception of surrounding vehicles within the vehicle-mounted camera field of view. Basically, the majority of car detection works are focused on appearance-based features [26], ranging from low-level features, e.g., edge, symmetry, HOG detectors [27]–[29] to high-level features such as those generated from deep convolutional neural networks and vehicle part-based SVM models [30], [31]. It is interesting to note that the state-of-the-art schemes are able to robustly localize the preceding and oncoming vehicles, even in the presence of vehicle part occlusions.

However, the pure vision system has its own limitations. It suffers from object scale ambiguity and global orientation drifts. To overcome these problems, Dong *et al.* [32] have exploited both inertial and visual cues to analyze the vehicle and its pose semantically. The semantic (identity) and syntactic (poses) attributes of vehicles are decomposed into a geometric localization-and-mapping term and the likelihood function model approximated by pretrained convolutional neural networks. But the vehicle analysis is still limited to static scenes, due to the lack of sophisticated class-specific deformation prior inference. It is known that IMU belongs to the proprioceptive sensor type; while by contrast, visual system belongs to the exteroceptive sensor type. IMU is able to help visual system in distinguishing the moving scenes and static scenes. The static scenes could be exploited to compute the vehicle motions; the moving scenes could be utilized to analyze the surrounding vehicles. However, few works have considered this visual-inertial complementary property.

III. PROBLEM DESCRIPTION

Basically, the detected feature correspondences can be divided into three categories: static inliers, dynamic inliers and outliers. Static inliers represent the correct feature matches that lie on static object regions, for instance, buildings, sidewalks, traffic signs and trees. They can be used to estimate the vehicle dynamics. By contrast, dynamic inliers represent the correct feature matches that lie on surrounding object regions that are moving, such as walking pedestrians,

running motorcycles and moving cars. They can be applied to analyze the dynamic traffic scenes. Outliers represent the incorrect feature matches due to similar blob patterns, occlusions, etc.

One of the fundamentals in vehicle dynamic analysis and surrounding environment analysis is to accurately identify static inliers, dynamic inliers and outliers. Besides, effective detection and localization of moving vehicles also play an important role in surrounding environment analysis. Thus, in this section, the problems in feature match classification and vehicle detection will be briefly described.

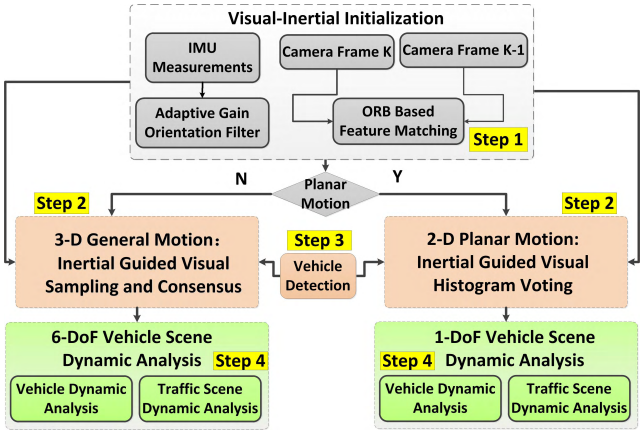


FIGURE 2. The framework for the proposed visual-inertial based vehicle and surrounding environment dynamic analysis.

The flowchart of the proposed approach is shown in Fig. 2. In the first step, the keypoint features between consecutive image frames are detected and matched using ORB feature detection and matching. It is noticeable that the detection scheme adopts a uniform keypoint distribution in the image by suppressing the weaker-response corners. Then these features are fed into the geometry constraint model of ground vehicles. Meanwhile, the inertial measurements: vehicle linear accelerations and angular velocity are preintegrated between the successive frames to suppress the inertial measurement errors [25]. In the second step, the instantaneous vehicle motion attribute (2-D planar motion or 3-D general motion) would be determined by fusing visual measurements with roll and pitch information derived from inertial measurements [33]. For the 3-D general motion, the inertial guided visual sampling and consensus model [34] is employed. For the 2-D planar motion, the inertial guided visual histogram voting is designed. (2-D planar vehicle motion analysis is also the main contribution of this paper). In the third step, the part based vehicle detection model is developed to explicitly identify the vehicle regions. Within each region, the features of moving vehicle are matched. Eventually, the vehicle dynamics are estimated using static inliers; the surrounding environment is analyzed through the detected vehicle regions and the corresponding dynamic inliers within. The estimated vehicle dynamics, in turn, able to compensate for the IMU long-term attitude drifts (which can be referred in our previous works [34]).

IV. INERTIAL GUIDED VISUAL FEATURE LABELING

In this section, the feature labeling for 2-D planar motion and 3-D general motion will be described respectively in detail. Then the feature correspondence classification will be presented.

A. FEATURE LABELING FOR 2D PLANAR MOTIONS

1) NONHOLONOMIC-CONSTRAINT MODEL

For a vehicle moving on the ground plane, it follows the nonholonomic constraint. Its Ackermann steering assumption is that the moving trajectory can be approximated by a series of circular and planar paths [12]. If the wheeled vehicle is not running on the planar surface (for example, the vehicle experiences the slippage on the icy road or runs across the speed bumps), the non-holonomic constraint model will be relaxed to the general 3D motion estimations, which will be discussed in the subsection **B: Feature Labeling for 3D Planar Motions**.

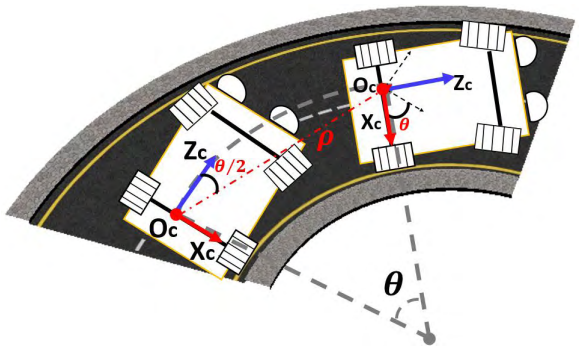


FIGURE 3. Vehicle nonholonomic-constraint geometry.

In this section, the non-holonomic constraint model will be described within the visual-inertial fusion framework. Suppose the camera frame origin O_c is rigidly fixed above the vehicle rear wheel axis. Its X-axis parallel to the back wheel axis; its Y-axis vertical to the ground; Z-axis perpendicular to the wheel axis (see Fig. 3). On the basis of camera nonholonomic constraint, the vehicle transformation $T = [R,\ t]$ can be recovered by:

$$t = \begin{bmatrix} \rho \cdot \sin \theta/2 \\ 0 \\ \rho \cdot \cos \theta/2 \end{bmatrix} \quad R = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix} \tag{1}$$

where t and R are the translational and rotational components respectively. θ denotes the rotational angle; ρ represents the translational length, which could be neglected since it is not exploited in the following derivations in separating the static inliers among the feature mathces.

Given a feature match $\{p,\ p'\}$ in camera coordinate between frame k and frame $k+1$, p and p' satisfy the epipolar constraint:

$$p'[t]_{\times}Rp = 0 \tag{2}$$

From the feature correspondence $\{p, p'\}$ and (2), the rotatory angle θ can be estimated as follows:

$$\theta = 2 \cdot \tan^{-1}\left(\frac{-p_z{}' p_x + p_x{}' p_z}{p_y{}' p_z + p_y p_z{}'}\right)$$
$$p = (p_x, p_y, p_z), \quad p' = \left(p_x{}', p_y{}', p_z{}'\right) \tag{3}$$

where (p_x, p_y, p_z) and $\left(p_x{}', p_y{}', p_z{}'\right)$ denote the coordinates of feature point p and p' in camera frame respectively. It is interesting that the moving agent pose could be parameterized by θ from the correct feature match (p_x, p_y, p_z) and $\left(p_x{}', p_y{}', p_z{}'\right)$ through (1).

2) INLIER/OUTLIER STATISTICAL MODELING

During vehicle dynamic analysis, the presence of incorrect data matches, among the putative correspondences, may lead to trajectory deviations at various levels. Worse still, it will accumulate over time and give rise to undesirable trajectories in a long term.

In order to solve this issue, the state-of-the-art works adopt One-Point-RANSAC or Histogram Voting to effectively identify and remove these outliers [12]. Since only one feature match is needed to minimally parameterize the motion model, the number of iterations is dramatically decreased. For instance, only seven iterations are required for One-point RANSAC given the probability success rate $p = 99\%$ and outlier percentage $\varepsilon = 50\%$ for $N = \frac{\log(1-p)}{\log\left(1-(1-\varepsilon)^S\right)}$. By comparison, the eight-point method requires at least 1177 iteration times to maximally remove the outliers and parameterize the motions.

Each feature match $\{p_i, p_i'\}$ is converted to rotation angle θ_i using (3). Then the histogram related to the group $\{\theta_i\}_{1\le i\le N}$ is built. In the histogram, every bin denotes the number of features that lie within the region $\theta_{i,1\le i\le N}$. Assuming that the group of inliers conforms to Gaussian Distribution $N(\mu, \sigma)$, where most of the inliers lie within $[\mu - 3\sigma, \mu + 3\sigma]$; the group of outliers conforms to Uniform Distribution $U(-180°, 180°)$.

However, this assumption does not hold in the scenario shown in Fig. 4 (c). Comparatively, the inlier group conforms to Multi-Gaussian Model $\left\{D_{j,0\le j\le M} \left| D_j = N\left(\mu_j, \sigma_j^2\right)\right.\right\}$ rather than the single Gaussian Model.

By using the Maximum A Posteriori (MAP) to model the inliers and outliers, the following can be obtained

$$\begin{aligned} &P_\theta\left(\left\{\left(p_i, p_i{}'\right)\right\}_{i=1}^N\right) \\ &= \prod_{i=1}^N P_\theta\left(p_i, p_i{}'\right) \\ &= \prod_{i=1}^N \prod_{j=0}^M \underbrace{\left\{P\left(\theta_i \middle| \theta_i \in N\left(\mu_j, \sigma_j^2\right)\right) P\left(\theta_i \in N\left(\mu_j, \sigma_j^2\right)\right)\right\}^{s(v_j)}}_{inlier} \\ &\quad \cdot \underbrace{\left\{P\left(\theta_i \left| \theta_i \in U\left(m, n\right)\right.\right) P\left(\theta_i \in U\left(m, n\right)\right)\right\}^{s(v_0)}}_{outlier} \end{aligned} \tag{4}$$

where $P\left(\theta_i \left| \theta_i \in N\left(\mu_j, \sigma_j^2\right)\right.\right)$ means the posteriori probability of θ_i lies in the j_{th} Gaussian clustering region; N is the number of feature matches; M is the number of inlier clusters, which is predetermined by the vehicle detection model introduced in Section V; It is noticeable that if there exists missed vehicle detection, the feature matches for these vehicle regions would be taken as static inlier $s(v_0)$ or outlier $s(v_o)$ (where subscript "o" is short for outlier).

$s(\cdot)$ is the "sign" function defined by

$$s\left(v_j\right) = \begin{cases} 1 & if\ \theta_i \in j_{th}\ inlier\ group \\ 0 & ohterwise \end{cases} \tag{5a}$$

$$s\left(v_o\right) = \begin{cases} 1 & if\ \theta_i \in outlier\ group \\ 0 & otherwise \end{cases} \tag{5b}$$

By taking the logarithm operation of (4) and removing the constant term, which is irrelevant to label assigning, it can obtain that

$$\begin{aligned} &\log\left(P_\theta\left(\left\{\left(p_i, p_i{}'\right)\right\}_{i=1}^N\right)\right) \\ &= \sum_{i=1}^N \sum_{j=0}^M \{s_{\theta_i}\left(v_j\right) \log\left(P\left(\theta_i \middle| \theta_i \in N\left(\mu_j, \sigma_j^2\right)\right)\right) \\ &\quad + s_{\theta_i}\left(v_o\right) \log\left(\theta_i \left| \theta_i \in U\left(m, n\right)\right.\right)\} \end{aligned} \tag{6}$$

The maximum likelihood in (6) is unable to be solved due to the unknown $s_{\theta_i}\left(v_j\right)$ and $s_{\theta_i}\left(v_o\right)$. This issue can be solved in a heuristic manner using expectation maximization (EM). EM proceeds iteratively within expectation step and maximization step:

- In the expectation step, the hidden variables, or say $s_{\theta_i}\left(v_j\right)$ and $s_{\theta_i}\left(v_o\right)$, are computed from the previous update θ_i: if $s_{\theta_i}\left(v_j\right) = 1$, θ_i belongs to the j_{th} inlier group; if $s_{\theta_i}\left(v_o\right) = 1$, θ_i belongs to the outlier group;
- In the maximization step, the conditional likelihood or say (6) is maximized, which is similar to estimate each Gaussian model mean value μ_j and covariance σ_j.

3) INERTIAL GUIDED VISUAL HISTOGRAM VOTING

For the 2-D planar motion, the vehicle moving parameter θ^* could be selected by choosing the mean value among the Gaussian Mixture Model that is consistent with the majority of samples. However, it fails if the majority of image areas are taken up by moving vehicles, in which case, most of correct matches may lie on moving vehicle regions rather than static object regions.

Fig. 4 is an example that shows the potential failure of static inlier selection using traditional methods. In Fig. 4 (a), the up-and-down image pair indicates the vehicle is making a right turn. The camera mounted onto the vehicle captures front-view scenes, which include both static and dynamic objects. Unfortunately, among the image-pair, most of feature matches are lying on the texture-rich truck regions, which are framed by red rectangles (see Fig. 4 (b), the feature matches in this region are connected by yellow lines). The rest of

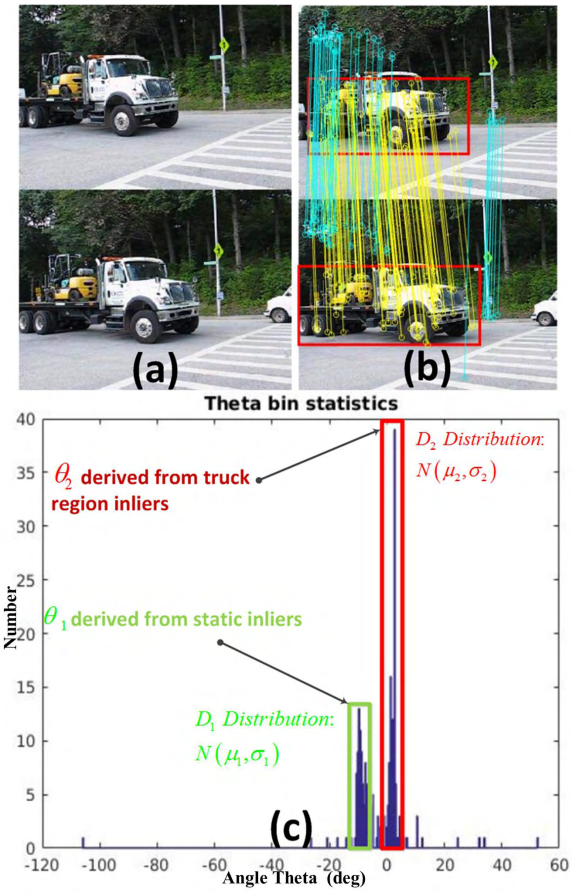


FIGURE 4. Illustration on failure of static inlier selection using traditional histogram voting. (a) original image pair; (b) vehicle detection and feature matching; (c) θ histogram voting.

detected feature matches lie on stationary areas: trees, traffic signs, street lamp and zebra-crossings.

In the case of Fig. 4(c), the static inliers conform to D_1 distribution; the inliers that lie on moving truck conform to D_2 distribution. It can be seen that the number of inliers that lie in D_2 area is larger than that in D_1 area. This enables the traditional histogram voting mistakenly choose D_2 as the ''inlier'' distribution; the estimated angle will be chosen as μ_2 rather than the correct μ_1.

In order to choose the most appropriate Gaussian model for the static inliers among the GMM, the rotatory angle θ_{imu} from inertial measurements is employed, to infer the inlier group that lie on static regions.

Suppose the inertial measurements consist of three-tuple $\{\alpha, \omega, m\}$. Vector α denotes tri-axial acceleration; vector ω denotes tri-axial angular rate; and vector m denotes tri-axial magnetic strength. They are jointly fed into the adaptive gain orientation filter (AGOF) [35] to derive the quaternion q_{imu} using

$$q_{imu} = k_t q_\alpha + (1 - k_t)\, k_t q_w \tag{7}$$

where q_α is the quaternion derived from accelerations and magnetic strengths; q_w is the quaternion derived from angular rates; k_t is the adaptive gain. Given $q_{imu} = (q_w, q_x, q_y, q_z)^T$, the corresponding rotational yaw angle θ_{imu} is derived as follows:

$$\theta_{imu} = \tan^{-1}\left(\frac{2 \cdot (q_y \cdot q_z + q_w \cdot q_x)}{q_w^2 + q_z^2 - q_x^2 - q_y^2}\right) \tag{8}$$

Suppose θ_{imu} from IMU conforms to Gaussian Distribution $N(\mu_{imu}, \sigma_{imu})$. The relative entropy [36] is measured between $N(\mu_{imu}, \sigma_{imu})$ and $N\left(\mu_j, \sigma_j^2\right)_{0\le j\le M}$ to infer θ^* as follows

$$D_{RE}\left(P_{imu} \| P_j \sim N\left(\mu_j, \sigma_j\right)\right) = \int_{-\infty}^{+\infty} p_{imu}(x) \ln \frac{p_{imu}(x)}{p_j(x)} dx \tag{9}$$

where $D_{RE}\left(P_{imu} \| P_j \sim N\left(\mu_j, \sigma_j\right)\right)$ is the relative entropy between the continuous random variable P_{imu} and P_j. P_{imu} represents the Gaussian distribution of IMU rotatory angle θ_{imu}; P_j represents the j_{th} Gaussian distribution of GMM. $p_{imu}(x)$ and $p_j(x)$ symbolize the probability density of P_{imu} and P_j respectfully. By choosing the minimum relative entropy among $\left\{D_{RE}\left(P_{imu} \middle\| P_j \sim N\left(\mu_j, \sigma_j^2\right)\right)\right\}_{0\le j\le M}$, the desirable vehicle ego rotation angle θ^* is obtained by

$$\theta^* = \arg\min_{\mu_j}\left\{D_{RE}\left(P_{imu} \middle\| P_j \sim N\left(\mu_j, \sigma_j^2\right)\right)\right\}_{0\le j\le M} \tag{10}$$

Given θ^*, it will be able to determine the Gaussian Model corresponding to θ^*. Then the static inlier group will be grouped.

B. FEATURE LABELING FOR 3D PLANAR MOTIONS

In case of vehicle moving in 3-D space, the Inertial Guided Visual SAmpling and Consensus (or say IGVSAC) is utilized to analyze the feature label. Assume that the visual putative correspondence set is $P_n = \{(p_1, p_1^{'}), (p_2, p_2^{'}), \cdots, (p_n, p_n^{'})\}$, where P_n contains both inliers and outliers. The variables to estimate are $O_n = \{R, T, S_n\}$. R and T represents the rotational matrix and translations respectively; They can be estimated through epipolar geometry constraint [34]. S_n consists of n feature labels (inlier or outlier), given by

$$S_n = (s_1, s_2, \cdots s_n) \tag{11}$$

where $s_{i,(1\le i\le n)} \in \{0, 1\}$ represents the i_{th} feature correspondence label, given as

$$s_i = \begin{cases} 1 & if\ s_i \in i_{th}\ inlier\ group \\ 0 & ohterwise \end{cases} \tag{12a}$$

$$s_o = \begin{cases} 1 & if\ s_i \in i_{th}\ outlier\ group \\ 0 & ohterwise \end{cases} \tag{12b}$$

The objective is to seek the model which is able to maximize $P(R, T)$ through the image potential feature matches and the initial rotational matrix R_0 obtained from inertial measurements.

Using Bayesian Rule, the goal in seeking the optimal $\emptyset_n^*$ can be obtained by maximizing the posterior probability

$$O_n^* = \arg\min_{O_n} P\{O_n|P_n\} = \arg\max_{O_n} \prod_{i=1}^{n} P\{s_i|\emptyset_n\} P(s_i|R,T) P(R,T) \quad (13)$$

The optimal solution on $\emptyset_n^*$ can be derived through Expectation Maximization (EM). For further details, please refer to our previous work [34].

C. INLIER/OUTLIER CLASSIFICATION

The static inliers are identified by using the orientations deduced from inertial measurements. The next step is to separate the dynamic inliers and outliers. Thus, this section will focus on the dynamic inlier classification. The dynamic inliers are identified in two steps. First, the car detection model is adopted to explicitly determine the number of vehicles as well as their corresponding vehicle regions in the camera field of view. Second, the vehicle regions between the image pair will be matched. Then each correspondence will be added to a specific vehicle label. For example, the feature match that lie in the q_{th} vehicle region, will be tagged label q. In terms of feature matches that lie in the non-vehicle regions, will be tagged label *void*. The group of feature matches are classified in this manner. For each vehicle region, there may exist both dynamic inliers and outliers. Separating them is also critical for accurately estimating surrounding environment. Here, the sample consensus is performed upon the feature matches that lie in each moving vehicle regions.

It is noticeable that if the numbers of detected vehicles are different between the adjacent frames or the vehicle detection fails in some specific vehicle regions, the feature matches lying within the unmatched vehicle regions will not be taken as dynamic inliers. Instead, these feature matches will be taken as static inliers or outliers by comparing the relative entropy with inertial orientations. Besides, during the vehicle region matching, the vehicle region correlations are adopted that consider both vehicle templates and nearest neighbor constraint, which considerably lowers the odds in incorrect vehicle matchings. In the presence of vehicle occlusions or missed detections, the feature pairings from these regions will be mistakenly labeled as outliers.

V. VEHICLE DETECTION AND LOCALIZATION

In this paper, the part-based model is used to detect vehicles in images. The part based model consists of a root filter and part filters. The root filter approximately localizes the entire vehicle regions; the part filter covers each ingredient of vehicle components, such as head & rear lamps, side doors, rear trunk, etc. The vehicle detection model includes a group of root-filter nonterminal units Ω and part-filter terminal units Γ. Both of them are organized through a hierarchical structure to describe the vehicle parts. The basic principle is to detect the potential vehicle regions that have high root-filter scores $\{\Omega_1, \Omega_2, \cdots, \Omega_R\}$. Then each potential region $\Omega_{c(1\leq c\leq R)}$ goes through the group of part filters. The responses of both the root filter and the part filters are added up to derive the final root location scores [37].

Contrary to the works in [31] and [37], the vehicle detection grammar is redefined to make the model suitable to the vehicle-mounted-camera configurations. It is common that the structure deformation frequently occurs in images due to camera view point changes. Thus, the vehicle detection model is trained with samples from various viewpoints, i.e., front view, side view and rear view. Furthermore, the vehicle invariant regional features such as head & rear lamps, license plates, trunk boot, side doors are chosen as invariable features to identify the vehicles.

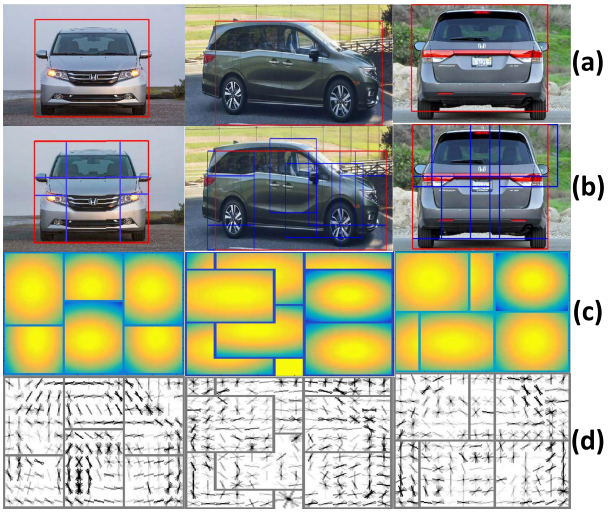


FIGURE 5. Vehicle detection model in front, side and rear views. (a) the original images with vehicles bounded by red rectangles; (b) vehicle components sub-bounded by blue rectangles; (c) the vehicle detector models; (d) the vehicle filter detectors.

The vehicle detection model is trained using the images from the proposed vehicle-mounted camera. Take the Honda Odessy vehicle detection as an example, the terminal and nonterminal regions are covered by the blue boxes and red box respectively shown in Fig. 5(b). The images are separated into front-view, side-view and rear-view classes. The model structures are trained separately. The vehicle detector structure models are displayed in Fig. 5(c). The vehicle detector that considers both the root filter and component filter is shown in Fig. 5(d). In this paper, the histogram of gradient features are chosen to describe the terminal components.

VI. VEHICLE AND SURROUNDING ENVIRONMENT DYNAMIC ANALYSIS

In this section, the visual static/dynamic inliers and inertial measurements are combined to estimate the vehicle dynamics and analyze the surrounding environment. The static inliers, together with preintegrated IMU states are jointly exploited within the nonlinear optimization framework. The procedure in estimating the surrounding environment are described afterwards.

A. VEHICLE DYNAMIC ANALYSIS

The keyframe based batch optimization is adopted to estimate vehicle states, because of its inherent advantage in repeated linearization and approximation of non-linear cost functions. The states include the 3 by 1 vehicle position vector: $P^W_{B,t}$, 3 by 3 rotational matrix: $R^W_{B,t}$, 3 by 1 speed vector: $v^W_{B,t}$ and 3 by 1 sensor bias vector: $b^B_{\alpha,t}$, $b^B_{\omega,t}$. The script denotes the state represented in which frame: body frame (local sensor frame) $\{B\}$ or world frame $\{W\}$.

The vehicle dynamics D_t, at time instant t, $D_t = \left\{P^{W\ T}_{B,t}, R^{W\ T}_{B,t}, v^{W\ T}_{B,t}, b^{B\ T}_{\alpha,t}, b^{B\ T}_{\omega,t}\right\}$ are estimated by jointly minimizing both the feature re-projection error $E_C(t,q)$ and inertial error $E_I(t,q)$ between the camera keyframe t and current frame q.

$$D^*_t = \arg\min_{\theta_t} \{s_c \cdot E_C(t,q) + s_t \cdot E_I(t,q)\} \tag{14}$$

where the coefficients s_c and s_i are the scale coefficients with regard to camera error term and IMU error term.

The camera feature re-projection error $E_C(t,q)$ between camera frame t and frame q is given by

$$E_C(t,q) = \sum_j \left\{\left(f_{t,j} - \pi\left(f_{q,j}\right)\right)^T \Sigma^q_{t,j} \left(f_{t,j} - \pi\left(f_{q,j}\right)\right)\right\} \tag{15}$$

where $f_{t,j}$ and $f_{q,j}$ are the j_{th} feature point locations at image t and image q respectively; $\pi(\cdot)$ denotes the feature re-projection at image domain; $\sum^q_{t,j}$ is the information matrix corresponding to the j_{th} feature correspondence. For further details in computing the re-projection errors and information matrix, please refer to [19].

The IMU error term $E_I(t,q)$ is given by

$$\begin{aligned} E_I(t,q) = &\left(e^P_{t,q}\right)^T \Sigma^P_{t,q} e^P_{t,q} + \left(e^v_{t,q}\right)^T \Sigma^v_{t,q} e^v_{t,q} + \left(e^R_{t,q}\right)^T \Sigma^R_{t,q} e^R_{t,q} \\ &+ \left(e^{b_\alpha}_{t,q}\right)^T \Sigma^{b_\alpha}_{t,q} e^{b_\alpha}_{t,q} + \left(e^{b_\omega}_{t,q}\right)^T \Sigma^{b_\omega}_{t,q} e^{b_\omega}_{t,q} \end{aligned} \tag{16}$$

where $e^P_{t,q}$, $e^v_{t,q}$ and $e^R_{t,q}$ are inertial position error, velocity error and rotation errors between time instant t and q respectively; $e^{b_\alpha}_{t,q}$ and $e^{b_\omega}_{t,q}$ are the acceleration bias and angular rate bias between t and q. $\sum^P_{t,q}$, $\sum^v_{t,q}$, $\sum^R_{t,q}$, $\sum^{b_\alpha}_{t,q}$ and $\sum^{b_\omega}_{t,q}$ denote the information matrix related to position, acceleration, rotation, acceleration biases and angular rate biases between time instant t and q, respectively.

These inertial error terms are given by

$$e^P_{t,q} = R^W_{B,t}\left(P^W_{B,t} - P^W_{B,q} - v^W_{B,t} \cdot \Delta t_{t,q} - \frac{1}{2} g \Delta t^2_{t,q}\right) - \Delta P^W_{t,q} \tag{17a}$$

$$e^v_{t,q} = R^W_{B,t}\left(v^W_{B,t} - v^W_{B,q} - g\Delta t_{t,q}\right) - \Delta v^W_{t,q} \tag{17b}$$

$$e^R_{t,q} = \log\left(\Delta R_{t,q} \cdot R^B_{t,t} \cdot R^B_{W,q}\right) \tag{17c}$$

$$e^{b_\alpha}_{t,q} = e^{b_\alpha}_t - e^{b_\alpha}_q \tag{17d}$$

$$e^{b_\omega}_{t,q} = e^{b_\omega}_t - e^{b_\omega}_q \tag{17e}$$

where g is the gravitational acceleration components; $\Delta t_{t,q}$ is the IMU sensing intervals between time instant t and q; $R^W_{B,t}$ and $P^W_{B,t}$ denote the rotation and translation of inertial frame with respect to world frame, respectively; $v^W_{B,t}$, $v^W_{B,q}$ represent the inertial velocity at time t and q in the world frame; $\Delta t_{t,q}$ is the time interval between t and q. $\Delta P^W_{t,q}$, $\Delta v^W_{t,q}$ and $\Delta R_{t,q}$ denote the expected translation, velocity difference and relative rotation between time instant t and q, respectively. $R^W_{B,t}$ is the rotation matrix of body frame with respect to world frame at time instant t and $R^B_{W,q}$ is the rotation matrix of world frame with respect to body frame (local sensor frame) at time instant q. $e^{b_\alpha}_t$ and $e^{b_\alpha}_q$ are the acceleration bias at time instant t and q respectively; $e^{b_\omega}_t$ and $e^{b_\omega}_q$ are the angular bias at time instant t and q respectively.

For more details in deriving the optimal vehicle dynamics using visual re-projection errors and inertial measurements, please refer to [19].

B. SURROUNDING ENVIRONMENT DYNAMIC ANALYSIS

In this subsection, the dynamics of surrounding vehicles captured by the camera, will be analyzed. Assume that the field-of-view moving objects are rigid. The non-rigidity like pedestrians are not taken into account in this work.

Suppose the vehicle regions in camera frame k is $\left\{V_{1,k}, V_{2,k}, \cdots, V_{M_k,k}\right\}$, and the number of vehicles displayed in frame k is M_k. Then the average number of vehicles M through the whole image sequence n would be

$$M = \frac{\sum_{k=1}^{n} M_k}{n} \tag{18}$$

Within each vehicle region, there exist both dynamic inliers and outliers. To separate the dynamic inliers, the histogram voting is performed upon these feature correspondences lying within the region, on the basis of which the dynamic inlier group $\left\{{}^{v_c}p_j, {}^{v_c}p'_j\right\}_{j=1}^{N_k(v_c,m)}$ that lies within the vehicle region V_c is obtained, where subscript j denotes the j_{th} dynamic inlier within the vehicle region; the superscript $N_k(v_c,m)$ denotes the number of dynamic inliers within the vehicle region v_c. Together with the number of static inliers $N_k(v_o)$, the average ratio of dynamic inliers R^k_{di} within the inlier group in the k_{th} image would be given by

$$R^k_{di} = \frac{\sum_{i=1}^{M} N_k(v_c,m)}{\sum_{i=1}^{M} N_k(v_c,m) + N_k(v_o)} \tag{19}$$

By means of the deduced parameters like the average number of vehicles, the average ratio of vehicle regions, the average ratio of dynamic inliers, the surrounding environment dynamics could be effectively analyzed, which could be potentially helpful during rush hours in downtown areas and interstate highways.

VII. EXPERIMENTS AND ANALYSIS

In this section, **V**ehicle and **S**urrounding **E**nvironment **D**ynamic **A**nalysis (VSEDA) will be extensively evaluated. At first, the vehicle-camera-IMU system will be briefly described. Then the experimental evaluations on VSEDA will be elaborated. The camera-IMU datasets were collected at various challenging traffic conditions, i.e., city urban environments and interstate highways. The datasets on urban environments were collected for four times (half of them were collected during rush hours and the other half during regular time). The number of captured images was more than 23,000 during rush hours and more than 10,000 during regular time. Likewise, the datasets on interstate highways were collected for four times (half of them were collected during rush hours and the other half were collected during regular time). The number of captured images was more than 12,000 during rush hours and more than 7,000 during regular time.

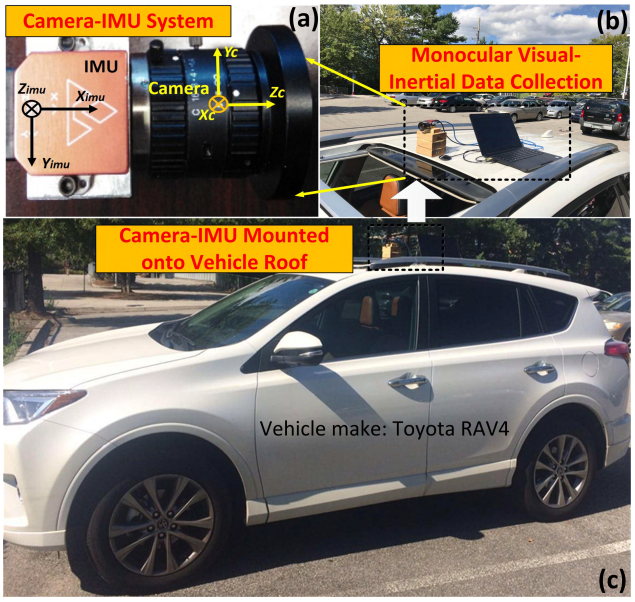


FIGURE 6. System configuration. (a) camera-IMU system; (b) the top view of camera-IMU system; (c) camera-IMU modules rigidly mounted onto the car roof.

A. SYSTEM SETUP

The experimental platform includes a monocular camera and an IMU. They were rigidly mounted on the top of the vehicle (Toyota RAV4) roof, as shown in Fig. 6. The camera ($XimeaMQ013CG-ON$) collected the data at $20fps$. The image resolution was 1280×1024 in pixels. The images have been corrected for radial distortion before the further processing. The IMU ($VectorNavVN-100$) collected the inertial measurements at $40Hz$. The extrinsics (translation and rotation of camera frame with respect to IMU frame) were calibrated using [38]. The intrinsic parameters on IMU biases and noise terms were calibrated using [39]. The IMU long-term drifts are compensated using the feedback control strategy [34]. The camera intrinsic parameters were estimated using the matlab calibration toolbox. The visual and inertial measurements were synchronized through timestamps. VSEDA was performed on the desktop with i7 processor and 8G memory. The DPM based vehicle detection could be processed within 40ms with GPU accelerations, which is able to achieve the real-time performance.

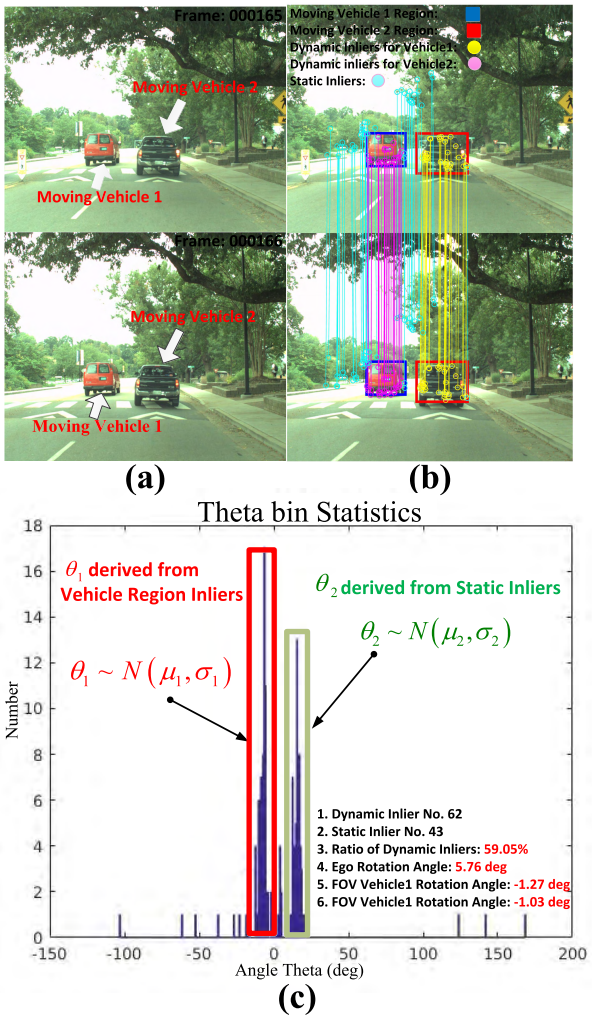


FIGURE 7. Example on inertial assisted histogram voting for VSEDA. (a) original vehicle successive images (b) framed vehicle regions and dynamic & static inliers matching; (c) inertial assisted histogram voting.

B. EXPERIMENTAL EVALUATION

For each image, the inertial cues were exploited firstly to check if the vehicle is moving on the planar ground. In the non-planar case, e.g., the vehicle went across the speed bumps, pits and steep slopes, the normal 6-DOF vehicle scene dynamic analysis was performed. For the planar ground scenarios, the 1-DOF vehicle scene dynamic analysis was conducted. It was found that for the whole trajectory, more than 70% of the ground were planar, so the 1-DOF VSEDA model worked for most of the time. In other words, the inertial guided histograming was suitable in most cases. One typical example in selecting the dynamic and static inliers is shown in Fig. 7. The number of dynamic inliers and

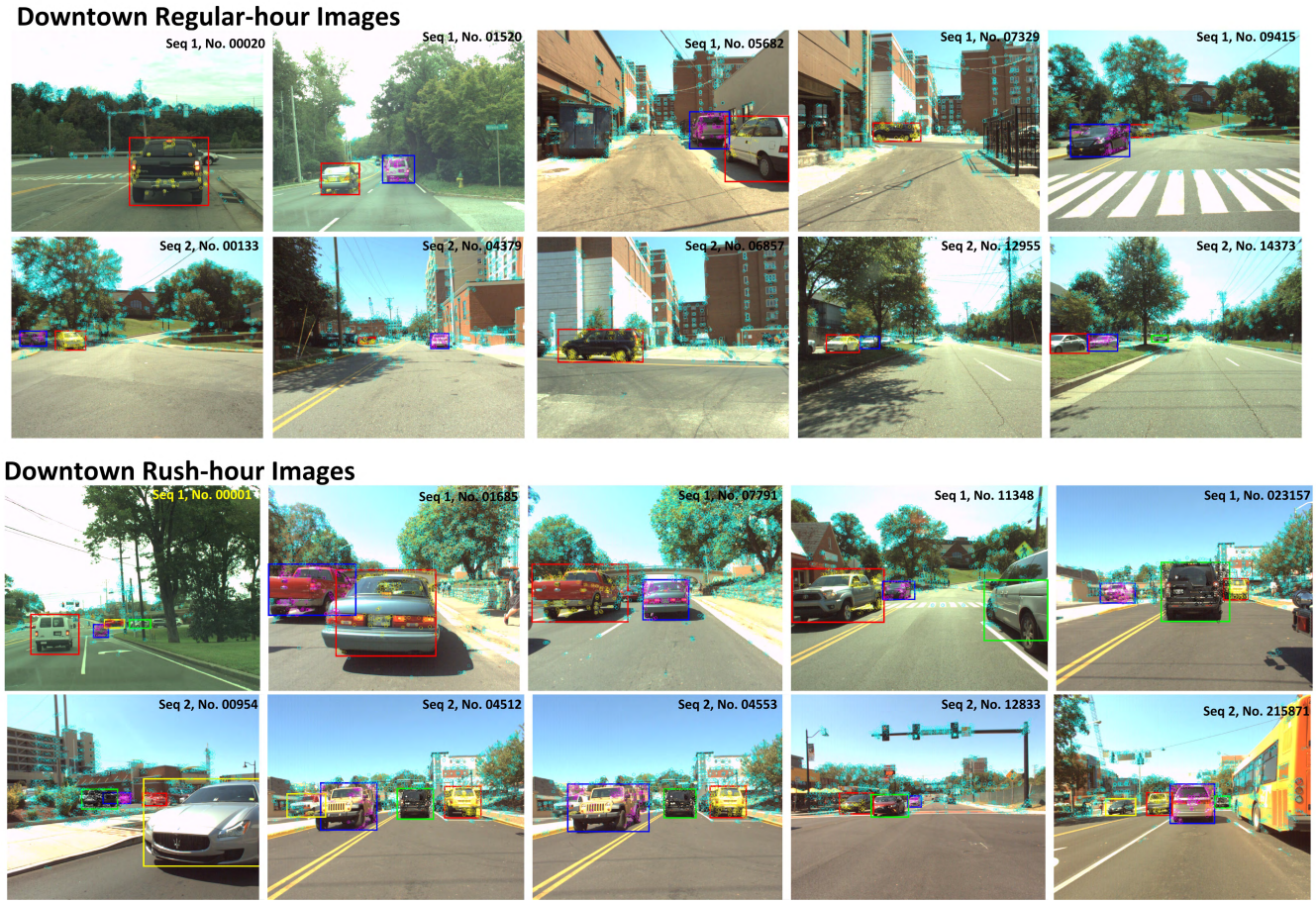


FIGURE 8. **Images collected in downtown areas during regular hours and rush hours.**

static inliers was 62 and 43 respectively. For the traditional majority voting, the θ represented by red boxes would be taken as "static" θ, which is undesirable for VSEDA. Using the proposed approach, by comparing the relative entropy between the inertial cues and Gaussian Mixture Model of θ regions, static inliers (represented by cyan circles) and dynamic inliers (represented by red and pink circles) could be distinguished among the feature matches. Then the vehicle dynamic analysis was performed using these static inliers; the scene dynamic analysis was conducted using these dynamic inliers, shown in Fig. 7(c).

In order to evaluate the proposed method for vehicle and surrounding environment dynamic analysis, we have collected the camera-IMU data at various challenging traffic conditions: during the rush hour and regular hour, in downtown areas and interstate highways. The camera-IMU data were stored in the laptop. Basically, one gigabyte of data were stored every minute.

1) DOWNTOWN AREA VEHICLE SCENE DYNAMIC ANALYSIS

In the downtown areas, especially during the rush hour, the traffic condition is relatively complex, i.e., the street is often jammed with a large number of vehicles and pedestrians walking on the sidewalk. Thus, vehicle surrounding environment dynamic analysis would be more meaningful in this scenario. Several representative images collected in downtown areas are shown in Fig. 8. It is noticeable that in some images, some vehicles were failed to be detected, e.g., in Fig. 8 downtown rush-hour image No. 215871. The orange-red campus bus was not detected. Yet, by means of inertial assisted feature histogram voting, our method is still able to distinguish the static inliers in the bus vehicle regions represented by the cyan color.

Fig. 9 shows two sample scenarios and the corresponding VSEDA results. Fig. 9 (a) and (c) were the images collected in the regular hour and rush hour respectively. It can be seen that the regions of vehicles, were explicitly detected and localized; the static and dynamics inliers were correctly identified. It should be noted that in Fig. 9 (c), the inliers represented by yellow color in the red frame still belong to the static type, since the corresponding vehicle was being parked on the parking lot.

Table 1 and 2 give the vehicle detection performance analysis and comparisons upon the collected challenging datasets in downtown areas. The data were collected two times. The total trajectory was approximately 3.25 kilometers. It took

FIGURE 9. Downtown area vehicle and traffic scene dynamic analysis.

TABLE 1. Downtown area vehicle detection performance analysis.

Downtown Area	Dataset 1		Dataset 2	
	Rush Hour	Regular	Rush Hour	Regular
Precision	99.59%	99.83%	99.27%	99.85%
Recall	98.23%	99.36%	97.95%	99.48%
F-measure	97.88%	99.59%	98.61%	99.66%

TABLE 2. Downtown area vehicle detection comparisons.

Downtown Area	Precision	False alarm	Miss rate
Ours	99.49%	0.51%	0.06%
Method [40]	98.48%	1.34%	0.49%

about 20 minutes during the rush hour and 10 minutes during the regular time. In Table 1, the *precision*, *recall rate* and $F-measure$ were computed upon the datasets recorded in both rush hours and regular times. Comparably, the precision is relatively higher, reaching up to 99%. The *recall rate* is relatively lower, since some vehicles failed to be detected in presence of part occlusions. Moreover, the comparisons between the proposed DPM model and the symmetrical attribute model [28] are also compared upon the selected images among the dataset, shown in Table 2. It can be seen that the DPM model outperforms the method [28] in terms of precision, false alarm and miss rate indexes.

In Table 3, the re-projection errors are exploited to analyze the vehicle dynamics; the number of frames, number of vehicles and its precision, vehicle region ratios and its precision, dynamic inlier ratios were considered, to analyze the surrounding environment. It is noticeable that during the rush hour, the dynamic inlier ratios were relatively higher, e.g., the maximum ratios of dynamic inliers reached up to 80%. In this case, the traditional histogram voting methods would fail to identify the correct rotation angle θ.

TABLE 3. Downtown area vehicle and surrounding environment dynamic analysis.

Downtown Area	Dataset 1		Dataset 2	
	Rush Hour	Regular	Rush Hour	Regular
No. of frames	23721	11543	27854	14692
Ave. Re-projection errors (pixel)	0.791	0.257	0.677	0.128
Ave. No. of vehicles	3.97	1.26	3.85	1.43
Ave. No. of vehicle detection error	0.49%	0.25%	0.33%	0.21%
Ave. ratio of vehicle regions	27.78%	11.49%	23.58%	9.77%
Ave. ratio of vehicle region precision	99.13%	99.49%	99.22%	99.73%
Max. ratio of vehicle regions	33.45%	17.96%	31.22%	16.82%
Max. ratio of vehicle region precision	87.26%	92.69%	77.39%	96.42%
Ave. ratio of dynamic inliers	31.49%	14.53%	28.72%	11.29%
Max. ratio of dynamic inliers	72.98%	35.47%	81.93%	22.85%

However, the proposed method is able to effectively separate the dynamic inliers and static inliers, thanks to the assistance of inertial quaternions. Besides, the group of dynamic inliers corresponding to moving pedestrians, especially during the rush hours, also deteriorates upon the estimations, with re-projection errors 0.791 and 0.677 pixels, for the rush-hour tests. Compared with ground truth, the average number of vehicle detection precision and average ratio of vehicle region precision reaches up to 90%. However, when it comes to maximum ratio of vehicle region precision in the times of rush hours, it was only around 80%, which can be attributed to the vehicle detection failures in presence of large areas of vehicle occlusions.

Table 4 shows the comparisons of re-projection errors on downtown datasets using the proposed VSEDA and one-point ransac [12]. The re-projection errors were computed from the group of static inliers using epipolar constraint and Sampson error formula [40]. It should be noticed that, in [12], the omni-directional camera was configured on the vehicle roof, and there are more static inliers since the camera lens features in wider angles. By comparison, in our case, the forward-facing camera was fixed on the vehicle roof and captured

TABLE 4. Static Inlier Re-projection error comparisons on downtown datasets.

Downtown Area	Re-projections Error (pixels)			
	Regular1	Regular2	Rushhour1	Rushhour2
Voting [12]	3.57	4.28	8.66	10.25
Ours	0.13	0.26	0.68	0.79

the images with relatively narrower angle lens during rush hours in downtown areas. The comparisons between [12] and ours were performed in the camera forward-facing camera scenario. It can be seen that RANSAC yielded the larger re-projection errors, for the failure in separating dynamic inliers and static inliers. By comparison, the proposed approach yielded the smaller re-projection errors than RANSAC, since it is capable of separating the dynamic and static inliers in both 2D and 3D cases.

TABLE 5. Re-projection error comparisons for the inliers within vehicle regions on downtown datasets.

Vehicle Detection		Re-projection Error (pixels)			
		Regular1	Regular2	Rushhour1	Rushhour2
Detected	Parking	0.21	0.28	0.74	0.85
	Moving	0.92	0.86	1.83	2.17
Undetected	Parking	0.35	0.27	0.71	0.77
	Moving	5.13	6.42	9.39	10.26

Table 5 shows the re-projection errors of the inliers lying within the vehicle regions. It can be seen that if the surrounding vehicle is detected, the re-projection errors for the parking scenes are relatively lower than the errors for the moving scenes, especially during the rush-hour datasets. By comparison, if the surrounding vehicle is undetected and moving, the re-projection performance deteriorates considerably, with the errors more than 5 pixels and 9 pixels for the regular-hour and rush-hour datasets, respectively. The reason for this degradation is the failure in detecting the moving vehicles, which results in the inaccurate epipolar line for the dynamic inliers lying within the undetected vehicle regions.

2) INTERSTATE HIGHWAY VEHICLE SCENE DYNAMIC ANALYSIS

The vehicle and surrounding environment dynamic analysis was also performed upon the visual-inertial dataset collected on the interstate highway during both rush hours and regular hours. The vehicle speed ranged from 70 km/h to 100 km/h and the camera refreshing rate was able to maintain the collected image not blurred. The representative images are shown in Fig. 10. In the rush hours, more than 12,000 images were collected for two times; in the regular hours, more than 7,000 images were collected for two times. In comparison with the images collected in the downtown areas, the vehicle region ratio on the highway is relatively smaller. However, the bundle of features are more liable to lie on moving vehicle regions. Thus, to identify the dynamic and static inliers is still important.

Fig. 11 shows two sample scenarios and the corresponding VSEDA results. As can be seen in Fig. 11 (a) and (b), two vehicles and the corresponding dynamic inliers were detected and matched. The rotation angle θ deduced from the group of static inliers, was 0.7°. If the histogram voting is conducted without vehicle detections and dynamic inlier separations, θ would be 4.24°, which is far from the expectation. Likewise, Fig. 11 (c) and (d) show the VSEDA results in highway rush-hour situations. Four vehicles were detected and the vehicle region ratio was only 4.31%. In contrast, the ratio of dynamic inliers was as high as 59%, which means the majority of inliers lie on moving vehicle regions. If the VSEDA was not performed, histogram voting would yield θ as 6.75°. In fact, θ was only 2.06° in this scenario.

TABLE 6. Interstate highway vehicle detection performance analysis.

Interstate Highway	Dataset 1		Dataset 2	
	Rush Hour	Regular	Rush Hour	Regular
Precision	99.76%	99.97%	99.82%	99.88%
Recall	96.45%	98.76%	95.38%	99.22%
F-measure	98.08%	99.36%	97.55%	99.55%

TABLE 7. Interstate highway vehicle detection comparisons.

Downtown Area	Precision	False alarm	Miss rate
Ours	99.22%	0.79%	0.04%
Method [40]	98.11%	1.52%	0.33%

Table 6 and 7 depict the vehicle detection performance and comparisons for interstate highway scenarios. The total trajectory was approximately 21.24 kilometers. It took more than 10 minutes to sample the data during rush hours; and more than 7 minutes to sample the data during regular hours. In Table 6, the precision is still able to reach up to 99%. But the recall rate is relatively lower, especially during the rush hours, which can be attributed to the fact that the distance between the vehicle and the surrounding moving objects are longer and the vehicle part based model fails sometimes if the features are not that obvious. In Table 7, our detection method outperforms the method [28] in terms of precision, false alarm and miss rate, upon the representative images selected among the datasets.

As can be seen in Table 8, compared with the visual-inertial data for downtown areas, the ratios of vehicle regions, in highway image datasets, were relatively smaller. This is mainly because the driver is required to keep a longer distance from vehicles ahead. Thus, the feature matches are more likely to be in the static regions. Besides, there are no pedestrians walking on the interstate highway, which makes the identified static inliers more reliable. These two factors enable the re-projection errors relatively smaller than those in downtown areas. Also, by comparison with ground truth, the average

FIGURE 10. Images collected on interstate highways during regular hours and rush hours.

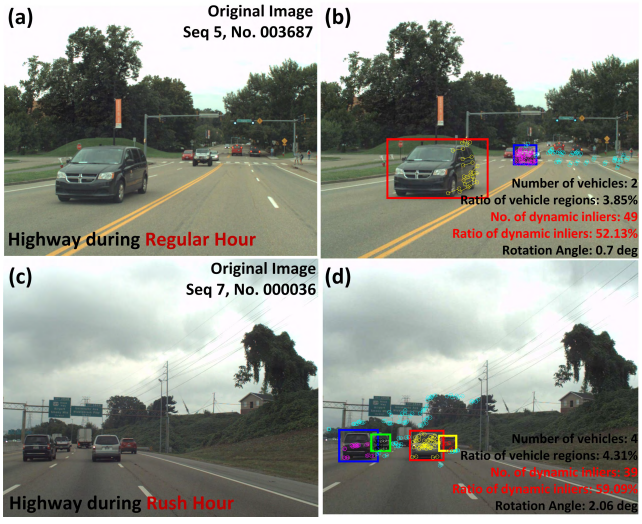


FIGURE 11. Interstate highway vehicle and traffic scene dynamic analysis.

number of vehicle detection precision and average ratio of vehicle region precision reaches up to 90%. However, when it comes to maximum ratio of vehicle region precision in the times of rush hours, it was less than 70%, which can be attributed to the fact that there were some scenes with multiple cars being exhibited from the car dealer. The car regions are small and unable to be distinguished from a long distance using our detection method.

From Table 9, it can be seen that both the proposed approach and other two state-of-the-arts yielded larger re-projection errors of inliers in rush hours than those in regular hours, which is mainly because there are more moving vehicles appearing in camera field of view. RANSAC [12] yielded the largest re-projection errors, approximately 10 pixels. By comparison, the proposed approach considers both planar & circular motion and general 6 DOF motions. In the general 6 DOF case, the inertial guided visual sample and consensus will be chosen to select the static and dynamic inliers; in the planar ground case, the inertial guided visual histogram voting fashion would be chosen. Through this way, the re-projection error is reduced by one order of magnitude, which was as low as 0.5 pixels.

Table 10 shows the re-projection errors of the inliers lying within the vehicle regions from the interstate highway datasets. It can be seen that if the surrounding vehicle is detected, the re-projection errors for the parking scenes are relatively lower than those for the moving scenes, especially during the rush-hour datasets. If the surrounding vehicle

TABLE 8. Interstate highway vehicle and surrounding environment dynamic analysis.

Interstate Highway	Dataset 1		Dataset 2	
	Rush Hour	Regular	Rush Hour	Regular
No. of frames	12429	7807	14563	8457
Ave. Re-projection errors (pixel)	0.428	0.134	0.651	0.186
Ave. No. of vehicles	3.67	1.35	4.22	1.77
Ave. No. of vehicle detection error	0.24%	0.03%	0.18%	0.12%
Ave. ratio of vehicle regions	10.23%	2.24%	8.77%	1.75%
Ave. ratio of vehicle region precision	99.42%	99.65%	99.39%	99.87%
Max. ratio of vehicle regions	19.21%	5.62%	17.65%	7.93%
Max. ratio of vehicle region precision	75.64%	93.55%	68.39%	98.29%
Ave. ratio of dynamic inliers	26.21%	11.22%	22.91%	10.83%
Max. ratio of dynamic inliers	61.37%	28.95%	52.44%	29.29%

TABLE 9. Static inlier Re-projection error comparisons on interstate highway datasets.

Interstate Highway	Re-projections Error (pixels)			
	Regular1	Regular2	Rushhour1	Rushhour2
Voting [12]	1.35	1.92	11.48	9.65
Ours	0.13	0.19	0.43	0.65

TABLE 10. Re-projection error comparisons for the inliers within vehicle regions on downtown datasets.

Vehicle Detection		Re-projection Error (pixels)			
		Regular1	Regular2	Rushhour1	Rushhour2
Detected	Parking	0.16	0.18	0.64	0.75
	Moving	1.12	1.33	1.97	2.08
Undetected	Parking	0.21	0.22	0.83	0.76
	Moving	4.45	5.39	8.28	15.46

is moving and undetected, the re-projection performance also deteriorates considerably, with the errors more than 4 pixels for the regular-hour and and 8 pixels for rush-hour datasets, respectively. It is interesting to note that, in the rush-hour2 datasets, the re-projection error, for the vehicle moving and undetected case, is as high as 15.46 pixels. The reason is that there was a large school bus moving ahead and being occluded for a period of three minutes and our proposed part-based model failed to detect that vehicle during the run. So the dynamic inliers within this school bus region resulted in poor re-projection performance. By and large, for the interstate highway datasets, the ratio of the vehicle areas are relatively smaller than that in the downtown areas, and the re-projection errors are relatively smaller for the highway datasets.

VIII. CONCLUSION AND FUTURE WORK

In this paper, a novel framework for vehicle and surrounding environment dynamic analysis is proposed. It fully exploits the monocular visual and inertial complementary properties to analyze both the vehicle dynamics and surrounding scenes. For vehicle dynamic analysis, the static inliers are selected by separating the features consistent with inertial quaternions. For surrounding environment analysis, the vehicle regions are detected by the part-based detetion model and the corresponding dynamic inliers are explicitly detected within each detected vehicle regions. Finally, the vehicle dynamics is estimated using the static inliers; the surrounding environments are analyzed using dynamic inliers. The challenging visual-inertial datasets were collected in both rush hour and regular hour time, in downtown areas and interstate highways. The experimental results and analysis prove that the proposed monocular visual-inertial system is quite suitable to analyze the ground vehicle dynamics and surrounding environments, which could be potentially applied for assistive drivings.

In the future, the deep neural networks, such as the DetectorNet [41], [42], will be trained and applied to detect the vehicles in real-time scenarios. In the future, we will also try to design a more comprehensive moving object detectors to separate the dynamic inliers and to localize the moving vehicle regions, on a variety of datasets such as KITTI [43] and BDD [44].

ACKNOWLEDGMENT

The authors would like to thank Zhe Su and Yan Li for their heartful assistance in collecting the visual-inertial datasets.

REFERENCES

[1] A. Geiger, M. Lauer, C. Wojek, C. Stiller, and R. Urtasun, "3D traffic scene understanding from movable platforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 1012–1025, May 2014.

[2] J. Wei, J. M. Snider, J. Kim, J. M. Dolan, R. Rajkumar, and B. Litkouhi, "Towards a viable autonomous driving research platform," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2013, pp. 763–770.

[3] K. Ransikarbum, N. Kim, S. Ha, R. A. Wysk, and L. Rothrock, "A highway-driving system design viewpoint using an agent-based modeling of an affordance-based finite state automata," *IEEE Access*, vol. 6, pp. 2193–2205, 2018.

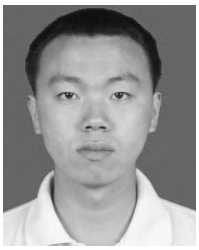
[4] L. Liu, H. Li, Y. Dai, and Q. Pan, "Robust and efficient relative pose with a multi-camera system for autonomous driving in highly dynamic environments," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2432–2444, Aug. 2018.

[5] Z. Duan, Y. Yang, K. Zhang, Y. Ni, and S. Bajgain, "Improved deep hybrid networks for urban traffic flow prediction using trajectory data," *IEEE Access*, vol. 6, pp. 31820–31827, 2018.

[6] A. Almagambetov, S. Velipasalar, and A. Baitassova, "Mobile standards-based traffic light detection in assistive devices for individuals with color-vision deficiency," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1305–1320, Jun. 2015.

[7] V. A. Shia *et al.*, "Semiautonomous vehicular control using driver modeling," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2696–2709, Dec. 2014.

[8] Y. Yang, L. Wu, S. Huang, J. Sun, W. Wan, and J. Wu, "Compensation details-based injection model for remote sensing image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 734–738, May 2018.
[9] A. Bar Hillel, R. Lerner, D. Levi, and G. Raz, "Recent progress in road and lane detection: A survey," *Mach. Vis. Appl.*, vol. 25, no. 3, pp. 727–745, 2014.
[10] R. Sabzevari and D. Scaramuzza, "Multi-body motion estimation from monocular vehicle-mounted cameras," *IEEE Trans. Robot.*, vol. 32, no. 3, pp. 638–651, Jun. 2016.
[11] A. Jaegle, S. Phillips, and K. Daniilidis, "Fast, robust, continuous monocular egomotion computation," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2016, pp. 773–780.
[12] D. Scaramuzza, "1-Point-RANSAC structure from motion for vehicle-mounted cameras by exploiting non-holonomic constraints," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 74–85, 2011.
[13] M. Lee, K.-S. Kim, and S. Kim, "Measuring vehicle velocity in real time using modulated motion blur of camera image data," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3659–3673, May 2017.
[14] D. Yadron and D. Tynan, "Tesla driver dies in first fatal crash while using the autopilot mode," *Guardian*, 2016. [Online]. Available: https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk
[15] H. Aliakbarpour, V. B. S. Prasath, K. Palaniappan, G. Seetharaman, and J. Dias, "Heterogeneous multi-view information fusion: Review of 3-D reconstruction methods and a new registration with uncertainty modeling," *IEEE Access*, vol. 4, pp. 8264–8285, 2016.
[16] G. Panahandeh and M. Jansson, "Vision-aided inertial navigation based on ground plane feature detection," *IEEE/ASME Trans. Mechatronics*, vol. 19, no. 4, pp. 1206–1215, Aug. 2014.
[17] L. Kneip, M. Chli, and R. Y. Siegwart, "Robust real-time visual odometry with a single camera and an IMU," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.
[18] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proc. IEEE Int. Conf. Robot. Autom.*, Apr. 2007, pp. 3565–3572.
[19] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robot. Res.*, vol. 34, no. 3, pp. 314–334, 2015.
[20] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 834–849.
[21] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep./Oct. 2015, pp. 298–304.
[22] M. Hwangbo, J.-S. Kim, and T. Kanade, "Inertial-aided KLT feature tracking for a moving camera," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2009, pp. 1909–1916.
[23] D. Kurz and S. B. Himane, "Inertial sensor-aligned visual feature descriptors," in *Proc. IEEE/RSJ Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 161–166.
[24] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular SLAM with map reuse," *IEEE Robot. Aut. Lett.*, vol. 2, no. 2, pp. 796–803, Apr. 2017.
[25] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual–inertial odometry," *IEEE Trans. Robot.*, vol. 33, no. 1, pp. 1–21, Feb. 2017.
[26] S. Sivaraman and M. M. Trivedi, "Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1773–1795, Dec. 2013.
[27] S. Tuermer, F. Kurz, P. Reinartz, and U. Stilla, "Airborne vehicle detection in dense urban areas using HoG features and disparity maps," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 6, pp. 2327–2337, Dec. 2013.
[28] J.-W. Hsieh, L.-C. Chen, and D.-Y. Chen, "Symmetrical SURF and its applications to vehicle detection and vehicle make and model recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 6–20, Feb. 2014.
[29] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 5, pp. 694–711, May 2006.
[30] S. Yu, Y. Wu, W. Li, Z. Song, and W. Zeng, "A model for fine-grained vehicle classification based on deep learning," *Neurocomputing*, vol. 257, pp. 97–103, Sep. 2017.
[31] H. He, Z. Shao, and J. Tan, "Recognition of car makes and models from a single traffic-camera image," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3182–3192, Dec. 2015.
[32] J. Dong, X. Fei, and S. Soatto, "Visual-inertial-semantic scene representation for 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3567–3577.
[33] N. Nourani-Vatani, J. M. Roberts, and M. V. Srinivasan, "IMU aided 3D visual odometry for car-like vehicles," in *Proc. Australas. Conf. Robot. Autom.*, 2008, pp. 1–8.
[34] Y. Zhang, W. Liang, Y. Li, H. An, and J. Tan, "Robust orientation estimate via inertial guided visual sample consensus," *Pers. Ubiquitous Comput.*, vol. 22, no. 2, pp. 259–274, 2018.
[35] Y. Tian, H. Wei, and J. Tan, "An adaptive-gain complementary filter for real-time human motion tracking with MARG sensors in free-living environments," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 21, no. 2, pp. 254–264, Mar. 2013.
[36] M. Zarinbal, M. H. F. Zarandi, and I. B. Turksen, "Relative entropy fuzzy c-means clustering," *Inf. Sci.*, vol. 260, pp. 74–97, Mar. 2014.
[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[38] H. He, Y. Li, and J. Tan, "Rotational coordinate transformation for visual-inertial sensor fusion," in *Proc. Int. Conf. Soc. Robot.*, 2016, pp. 431–440.
[39] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 2013, pp. 1280–1286.
[40] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[41] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2553–2561.
[42] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D boxes as CNN input for improved fine-grained vehicle recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3006–3015.
[43] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.
[44] H. Xu, Y. Gao, F. Yu, and T. Darrell. (2017). "End-to-end learning of driving models from large-scale video datasets." [Online]. Available: https://arxiv.org/abs/1612.01079



YINLONG ZHANG (S'14) received the B.S. degree in automation engineering, in 2010. He is currently pursuing the Ph.D. degree from the Shenyang Institute of Automation, Chinese Academy of Sciences, China. He is currently a Visiting Student with the University of Tennessee, Knoxville, TN, USA, where he is also performing the research. His current research interests are visual-inertial tracking, image processing, autonomous driving, and multi-sensor fusions.

WEI LIANG (SM'18) received the Ph.D. degree in mechatronic engineering from the Shenyang Institute of Automation, Chinese Academy of Sciences, in 2002. She is currently a Full Professor with the Shenyang Institute of Automation, Chinese Academy of Sciences. Her research interests include industrial sensor networks, body area networks, signal processing, and information fusion.

HONGSHENG HE (M'10) received the Ph.D. degree in electrical and computer engineering from the National University of Singapore, Singapore, in 2012. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, USA. His research interests include intelligent robotics, machine learning, and computer vision.

JINDONG TAN (M'02) received the Ph.D. degree in electrical and computer engineering from Michigan State University, East Lansing, MI, USA, in 2002. He is currently a Professor with the Department of Mechanical, Aerospace and Biomedical Engineering, The University of Tennessee, Knoxville, USA. His research interests include mobile sensor networks, augmented reality and biomedical imaging, dietary assessment, and mobile manipulation. He is a member of ACM and Sigma Xi.