IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received November 8, 2018, accepted November 26, 2018, date of publication December 14, 2018,
date of current version January 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886824

Robust Stereo Visual Odometry Based on Probabilistic Decoupling Ego-Motion Estimation and 3D SSC

YAN WANG, HUI-QI MIAO, AND LEI GUO, (Senior Member, IEEE)
School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

Corresponding author: Yan Wang (18811409162miao@gmail.com)

This work was supported by the National Nature Science Foundation of China under Grant 61573019, Grant 61751302, Grant 61320106010, and Grant 61627810.

ABSTRACT The outliers caused by noise and mismatching severely restrict the precision of visual odometry. Moreover, the dynamic environment is also a crucial element that decreases the robustness of the systems. This paper presents a robust stereo visual odometry by decoupled ego-motion estimation based on probabilistic matches and rejecting the outliers of dynamic objects through motion segmentation. Fast ZNCC method, based on local sum table and partition upper bound schemes, is presented for selecting probabilistic matches while keeping run-time efficiency. The selection of multi-correspondences can avoid mismatching of corresponding points. In consideration of noise interference, the essential matrix is computed in a probabilistic framework to estimate the initial value of the rotation matrix without estimated depth errors involved. Then, in order to estimate pose robustly in dynamic environment, a modified sparse subspace clustering (SSC) method is discussed, which aims to cluster the tracked 3D points cloud to avoid errors caused by affine transformation. The non-negative constraint makes the method suitable for fast moving camera. The proposed 3D-SSC method removes the outliers belonging to dynamic objects effectively. Finally, the detected inliers and depths are employed to estimate the translation matrix and refine rotation matrix. The proposed method is evaluated on the KITTI benchmark and compared with the state-of-the-art methods. The results show that our method is more robust as it can detect outliers more accurately in dynamic environments and achieve higher precision in motion estimation.

INDEX TERMS Stereo visual odometry, probabilistic matches, decoupling estimation, 3D SSC.

I. INTRODUCTION

In recent years, visual odometry (VO) and visual SLAM have played an immensely important role on autonomous driving [1], [2]. With the abundant information provided via vision, the autonomous system can generate self-localization measurements. Most odometry methods preform registration between the current image and a previous reference, in which the estimated transformations between these images are assumed to originate from the camera motion [3]. However, almost all the techniques are built under the assumption of static environments, which usually cannot be satisfied in the real world. Dynamic objects which violate this assumption will seriously influence the precision of estimation.

Improving the performance of visual odometry in dynamic environments is an important and desirable problem, especially for vehicles. For cameras equipped on vehicles capturing dynamic scenes, both static and dynamic scene parts appear to be moving [4]. It is seldom the case that vehicles operate in strictly static environments. Therefore, the moving objects in environments will significantly impact the accuracy of estimation.

In addition, the ego-motion of mobile vehicles consists of the rotation R and the translation t. They are estimated and integrated together in most VO approaches, which are prone to drift. From an application perspective, the location information of the vehicle is crucial information supported by the odometry. In the KITTI [5] vision benchmark scoreboard, the translation error is regarded as the exclusive factor for ranking, and the rotation error is displayed for reference only. Nevertheless, rotation errors have greater influence than translation errors on final location during the cumulation of errors in dead reckoning process such as odometry. The translation is reliant on the depth in contrast to the rotation. Therefore, the operation that estimates rotation and translation separately will recover rotation with extra precision and induce high accuracy for location [6]. The VO methods that involve feature detection and matching process, however, are suffering from feature mismatching [7].

Hence, in this paper, we design a probabilistic decoupled framework based robust ego-motion estimation algorithm to estimate rotation, with translation calculated after dynamic objects rejected by motion segmentation method. A pair of stereo cameras are employed to capture images of scenes and infer the ego-motion. In particular, a set of correspondence points accompanying with the probabilities of matching for each corner are estimated through fast ZNCC method to maintain both accuracy and efficiency. The use of probabilistic multi-correspondences allows us to hold several match hypotheses for the essential matrix computation, which is an advantage when there are ambiguous matches. Furthermore, to avoid the interruptions of dynamic objects in environment, the 3D trajectories of tracked points along sequences are employed to directly infer the clustering of data. Since the Sparse Subspace Clustering (SSC) method performs motion segmentation well in Hopkins dataset [8], here we modify this method with 3D data structure as well as non-negative constraint to categorize trajectories at high speed. After outliers rejection through 3D-SSC, the filtered matching points are used to estimate translation. The algorithm is tested on the KITTI dataset. The experimental results demonstrate that our approach is able to exclude outliers of moving objects in high speed scenes and improve the motion estimation performance. The main contributions of our work are as follows:

1) A novel motion estimation method is devised, which estimates the essential matrix in probabilistic framework with noise and mismatching error being considered. Therefore, the precision of rotation is improved by the use of robust probabilistic framework and decoupling estimation for the essential matrix.

2) The SSC method is employed in 3D space to avoid error caused by affine transformation. In addition, a more restrict constraint for sparse self-expression is applied to adjust to fast moving camera circumstances.

The paper is organized as follows. In Section II, we review state-of-art visual odometry methods as well as several approaches which tackle dynamic environments. Section III describes the overall structure of our visual odometry algorithm. The details about selection of probabilistic matches and decoupling estimation of rotation are given first. Then, the SSC method combined with 3D points (3D SSC) is discussed in detail. In Section IV, we assess the performance of our method by comparing it with state-of-art works. Section V concludes our work and states future directions for research.

II. RELATED WORK

The estimation of the ego-motion, the position and orientation of the car is addressed with wheel encoders traditionally, which suffers from wheel slip in uneven terrain or adverse conditions and cannot recover from errors in the measurements. The visual odometry technique first appeared in [9] and then became popular as it is less affected by these conditions. After that, more and more researchers and groups have been focusing on the topic of self-localization of a system relying solely or mainly on visual data.

The essential part of any visual odometry is the robustness of estimation. Therefore, a broad variety of methods have been introduced to increase the accuracy of estimation [10], [11]. The most general way is to estimate the full six motion parameters in a Random Sample Consensus (RANSAC) [12] framework [13], [14], where the reliability of full motion hypothesis is related to the number of iterations and the error threshold. This means that it is difficult to lead to a correct result. Several best performing methods in KITTI dataset decouple the estimation of the rotation and the translation as there is a fundamental difference between their estimation. An initial rotation estimation is utilized to decouple the rotational and translational optical flow in [15]. The resulting characteristics are then used to exclude outliers. While in [16] the motion is separately estimated, where the rotation is estimated by five point method and the translation is estimated by three point method. A separation step of rotation and translation estimation is carried out [17] in the condition of known direction and homography relation. Such studies mention that the depth is only relative to translation, in contrast to rotation. Following this idea that the rotation is independent on the depth, we estimate the rotation component in advance without depth information involved to acquire higher accuracy.

The above mentioned methods directly use feature correspondences between images to estimate the ego-motion. However, the deterministic matching points may lead to errors due to too little texture, low quality of images and noise. Then researchers tend to employ probabilistic method to reduce ambiguity. Domke and Aloimonos [18] propose a method to compute correspondence probability distributions using Gabor filters which are tuned to different orientations and scales. They further establish a probabilistic framework to estimate epipolar geometry. However, in the presence of regularly repetitive texture, the responses of the Gabor filter are identical at multiple places and this would lead to ambiguous estimation of ego-motion. Work in [19] models the joint probability distributions related to the positions of corresponding features in different images by using the joint feature distribution to yield a distribution over all feasible ego-motions. However, given the variates in scene structure, the extracted feature correspondences between images is not always possible and prone to be mismatched. Therefore, the result of joint feature distribution probably is not true. The method proposed in [20] and [21] presents a structure from motion algorithm by using optical flow probability distributions calculated by gradient based method. This method assumes that the noise is Gaussian and the noise parameters are chosen empirically to compute optical flow from the gradient of images, which may easily violate the actual situation.

Without any restriction of errors, the zero-mean normalized cross correlation (ZNCC) is served for probability distribution computation to acquire more robust ego-motion estimation under adverse image conditions [22]. Although this algorithm provides more accurate correspondences for motion estimation, the practical usage is limited because of the computational burden of ZNCC. Our approach, inspired by [22], raises a fast ZNCC method to intensely reduce the computation redundancy and keep its high precision. This kind of correspondences will achieve a more accurate ego-motion estimation, particularly when image is ambiguous, under non-ideal conditions and containing many unreliable correspondences.

Despite the remarkable results in visual odometry, most approaches work on the assumption of static environments. Since the real world usually contains dynamic objects, current approaches are prone to failure due to false correspondences or occlusion of previously tracked features. Standard outliers rejection approach such as RANSAC only work well under the circumstance where static features are the majority. To improve odometry performance in dynamic environment, [23] develops an automatic self-supervised approach to learn the recognition of dynamic objects in the environments, which does not require any manual labeling. Reference [24] also employs an image segmentation classifier learned from hand labeling training examples. This kind of training method requires enough training data and invariable environment for estimation. Other approaches leverage external sensors such as an inertial measurement unit (IMU) to solve this problem [25], [26]. A background model estimated from the warped depth images is developed by [27] to subtract static background. The depth captured from RGB-D camera severely restricts its application to outdoor environment.

Multibody motion segmentation clustering is another kind of technique to handle motion detection problem in keypoint based VO. In recent decades, numerous works have been developed in multibody motion segmentation, such as Generalized Principle Component Analysis (GPCA) [28] and RANSAC-based motion segmentation. Elhamifar and Vidal [29] proposed a SSC algorithm supported by the notion of self-expressiveness property of the data and spectral clustering framework as inferring solution. In fact, SSC is proved to be the best performance motion segmentation method [8]. Afterwards, many studies have been proposed to modify the SSC method for higher precision. Reference [30] considers the problem of subspace clustering under noise. While [31] and [32] address the problems of incomplete data and large size data, respectively. From a different perspective, the affinity and segmentation framework are joint optimized in [33]. Nevertheless, the challenge in visual odometry application is to tackle the segmentation in the case of high speed cameras. In this paper, the 3D SSC method is developed to acquire more robust segmentation results when vehicles move quickly.

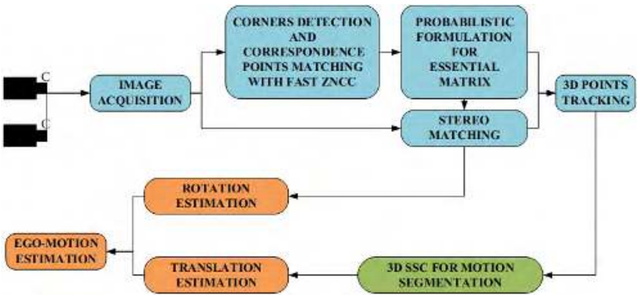

C
C
IMAGE ACQUISITION
CORNERS DETECTION AND CORRESPONDENCE POINTS MATCHING WITH FAST ZNCC
PROBABILISTIC FORMULATION FOR ESSENTIAL MATRIX
STEREO MATCHING
3D POINTS TRACKING
ROTATION ESTIMATION
EGO-MOTION ESTIMATION
TRANSLATION ESTIMATION
3D SSC FOR MOTION SEGMENTATION

FIGURE 1. The pipeline of our algorithm.

III. PROBLEM FRAMEWORK AND METHOD

This paper presents a visual odometry algorithm that uses not only probabilistic framework to achieve rotation decoupling estimation but also 3D SSC to detect outliers of the dynamic objects in high speed environment. As shown in the visual odometry pipeline of our proposed method in Fig. 1, the corresponding points of the current frame matching with each corner detected in the previous frame are computed via fast ZNCC method (see Section III-A). Using the probabilistic matches, the essential matrix is defined as the hypothesis with the maximum likelihood. Then the one-to-one correspondence is detected by epipolar constraint, which allows the 3D points to be tracked along the sequence (see Section III-B). To increase the robustness against dynamic environment, 3D SSC method is applied to distinguish the points between background and dynamic objects in direct 3D space (see Section III-C). Finally, the rotation extracted from essential matrix is refined by reprojection error and the translation is estimated more accurately after exclusion of outliers.

A. PROBABILISTIC CORRESPONDING POINTS MATCHING WITH FAST ZNCC

The first stage of the algorithm is to find the probabilistic corresponding points $q_{1,\cdots,m}$ in image I_{k+1} which match with corner p detected in I_k. For corner detection, we utilize Features from Accelerated Segment Test (FAST) algorithm [34] on the previous frame with non-maximum suppression employed. After the corners are detected, the bucketing scheme is applied to obtain uniformly distributed corners as described in [35] for following points clustering task. Then the most possible corresponding points and the related probabilities of each corner will be selected as input of ego-motion estimation in probabilistic framework. Here we utilize the ZNCC method for probabilistic corresponding points matching as [22]. The corresponding points in current frame of each corner is determined through ZNCC over a searching area around corner in previous frame. ZNCC is the most common and effective criterion of integer-pixel correlation calculation, since it is robust to changes in the amplitude of illumination on two compared images, and less sensitive to noise in comparison with the sum of absolute differences (SAD) and so on. However, there exists a higher computational cost because of its complex definition, which is a significant drawback in its real-time application.

In this section, we develop a fast way to derive ZNCC based corresponding points, while still keep robustness. For each corner, the computation of ZNCC values over the searching area can be split into numerator and denominator computation, respectively. The denominator is computed by local sum table to eliminate repetition and redundancy, while the numerator is computed with block partition and upper bound schemes to skip impossible matches and obtain the most likely corresponding points with high ZNCC values. The computation of denominator in ZNCC involves the calculation of the means and the squares of the searching areas in the latter image. Hence, the two sums of intensities in searching area can be efficiently computed by introducing the local sum table in columns, expressed as

$$\begin{aligned} S_g(u,v) &= \sum_{x\in W_x} I(x+u,v,t+dt) \\ S_{g^2}(u,v) &= \sum_{x\in W_x} I(x+u,v,t+dt)^2 \end{aligned} \tag{1}$$

to compute the sums in table at location (u, v), where W_x is the row size of patch around the feature point. In the local sum table approach, we compute the sum values of every column along the row direction and store it in the sum table in advance.

Referring to Cauchy-Schwarz inequality, the upper bound $\mu_0(u, v)$ of the cross correlation term $\psi(u, v)$ in numerator can be inferred as

$$\begin{aligned} &\psi(u,v) \\ &\quad= \sum_{x\in W_x}\sum_{y\in W_y}[I(x,y,t)\times I(x+u,y+v,t+dt)] \\ &\quad\leq \sqrt{\sum_{x\in W_x}\sum_{y\in W_y} I(x,y,t)^2}\sqrt{\sum_{x\in W_x}\sum_{y\in W_y} I(x+u,y+v,t+dt)^2} \\ &\quad= \mu_0(u,v) \end{aligned} \tag{2}$$

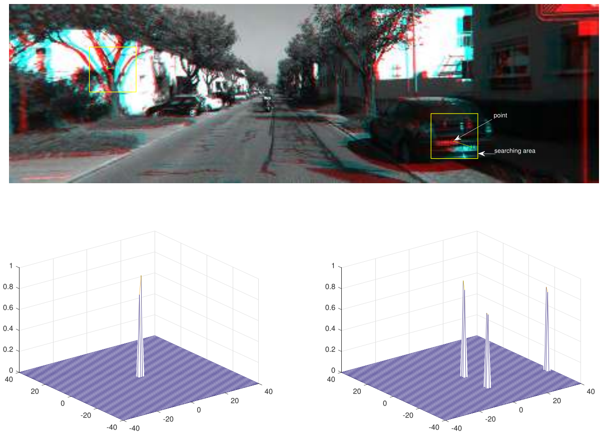
Following the multilevel successive elimination scheme, we divide the patch by column uniformly to obtain tighter upper bounds for different partitioning levels in cross correlation $\mu_l(u, v)$ as well as ZNCC value $C_{\max}$, respectively. The upper bound in l level is given as:

$$\begin{aligned} &\mu_l(u,v) \\ &= \sum_{y=1}^{l}\left(\sum_{x\in W_x}[I(x,y,t)\times I(x+u,y+v,t+dt)]\right) \\ &+ \sum_{y=l+1}^{W_y}\left(\sqrt{\sum_{x\in W_x} I^2(x,y,t)}\sqrt{\sum_{x\in W_x} I^2(x+u,y+v,t+dt)}\right) \end{aligned} \tag{3}$$

where W_x is the column size of patch and the maximum value of l. As the level number increases, the upper bound is more tighter. An initial value of maximum ZNCC value $C_{\max}$ is setted and updated, or the numerator computation is skipped as the upper bound C_μ of ZNCC value for certain level l is less than $C_{\max}$. The details of this algorithm are described in Algorithm 1.

Algorithm 1 Fast ZNCC Method Based on Local Sum Table and Partition Upper Bound Schemes

```
Require: Two images, the key points in the first image
Ensure: The final probability distributions of key points
1: Split the matching candidate patch in columns and establish the local intensity and square intensity sum tables in columns for each point.
2: Calculate the denominator by using sum table scheme to reduce repeating and redundant computations.
3: Start with calculating the numerator, initialize the level of the upper bound for cross correlation $\psi(u,v)$ as $l = 0$, where the number of column determines the hierarchy.
4: Caculate the upper bound of the level $l$.
   a) Calculate the cross correlations of the first $l$ parts.
   b) Obtain $\mu_l(u,v)$ by adding the result of (a) to the intensity squares of the last $N - l$ parts which are given by sum table $S_{g^2}(u,v)$.
   c) Derive the upper bound $C_\mu$ of the ZNCC value belonging to this level.
   d) Update $C_{\max}$ if $l = N$ or the current ZNCC value is bigger than $C_{\max}$.
5: Check if $C_\mu$ is larger than the current maximum ZNCC value. If not, set the ZNCC value in this position to 0 and go to step 1 to calculate the ZNCC value in the next position. Otherwise, go to step 6.
6: $l = l + 1$, go to step 4.
```



(a) (b)

FIGURE 2. The probability distributions of exemplary corresponding points with the partition upper bound computation.

The matching points and probabilities are illustrated in Fig. 2. It can be seen that only the most possible correspondence is left in Fig. 2(a), whereas Fig. 2(b) still remains several most likely correspondences with high probabilities. Other points in the searching area are skipped during the execution of upper bound scheme.

B. DECOUPLING ROTATION BY ESSENTIAL MATRIX ESTIMATION

As in vehicle tracking and locating, it is ineluctable in consecutive frames that a large number of mismatching pixels will be present. Therefore, the usage of multi-corresponding points of each corner for ego-motion estimation can avoid high mismatching errors and significantly improve the accuracy. The seminal work of Dmoke and Aloimonos [18] has dealt with the epipolar geometry estimation by computing the response of Gabor filters and further calculating the reliability of the essential matrix using an exponential transformation. In this section, the approach in [17] is briefly introduced firstly. Then the extension with perpendicular distance and Least-Median-of-Squares (LMedS) technique is carried out to improve the essential matrix estimation.

In [17], the parameters to be estimated are rotation matrix R and normalized translation directions $\hat{t} = \frac{t}{\|t\|}$. A bunch of hypotheses for parameters are defined and evaluated by analyzing the probabilistic correspondences $\rho_p(q)$ along the epipolar lines. A correspondence q for the corner p is selected which best satisfies the epipolar constraint denoted by

$$p^T Eq = 0 \tag{4}$$

For each point p in image I_k, the probability of one parameter hypothesis E_i is measured by the exponential probability of the optimal probabilistic corresponding point along the epipolar constraint in I_{k+1} , the probability is given by assuming each point maintains statistical independence of each other

$$\rho\left(E_i\right) \propto \prod \rho\left(E_i \left| p\right.\right) \tag{5}$$

Finally, the result of the algorithm is the parameters with the maximum probability as defined

$$E^* = \arg\max \rho(E_i^*) \tag{6}$$

Although the algorithm in [18] estimated the essential matrix in probabilistic framework, the error in epipolar geometric resulted from noise is not considered. Besides, simple multiplication is not very reliable because a certain probability is too large or too small with exponential operation. Our formulation makes significant revisions to the problem: the probability of a motion related to a point p is computed by combining the perpendicular distance from correspondence to the line Ep on image plane:

$$\rho\left(E_i \left| p\right.\right) \propto \max_{p^T Eq=0} \left(\rho\left(q\right) - \lambda d_{Ep\to q}\right) \tag{7}$$

where the parameter λ balances the two terms in computing probability. The correspondence is selected through fast ZNCC method. With considering perpendicular distance as error caused by noise and mismatching, the accuracy of rotation estimation can be less affected by outliers.

Furthermore, the probabilistic formulation combined with the LMedS [36] is used for calculating essential matrix. LMedS is one of the parameter estimation methods. It acquires the best model by minimizing the median deviation between samples and estimated model parameters. Hence, there is no need for LMedS to distinguish inliers and outliers with presetting threshold, compared with RANSAC.

Firstly, we randomly sample approximately 5000 points over the 5D space of essential matrix, and then the hypothesis which is mostly possible to be the correct solution is given by:

$$E^* = \arg\max_i \left\{ \operatorname{med} \left[\max_{p^T E_i q_j = 0} \left(\rho_{p_j}\left(q\right) - \lambda d_{E_i p_j \to q}\right) \right]_{j=1}^{n} \right\} \tag{8}$$

where n is the number of points detected in I_k. Our formulation then finds a median probability of each hypothesis among corners and chooses the one with the maximum probability. Finally, after some of the most probable motion parameters are selected, we employ the Nelder-Mead simplex algorithm to obtain the optimal motion parameters between two successive frames.

Now, the rotation matrix can be extracted from essential matrix without involving any depth information, which improves the accuracy of rotation between frames along with accumulated pose. The exact matches of corners in image I_k can now be searched through essential matrix in the epipolar line. And the disparities between $\left\{I_k^l, I_k^r\right\}$ and $\left\{I_{k+1}^l, I_{k+1}^r\right\}$ are computed to obtain the two sets of matched 3D points $\{P_k, P_{k+1}\}$ for translation estimation.

C. 3D POINTS BASED SSC FOR HIGH SPEED

The purpose of motion segmentation is to distinguish different motions between multi-trajectories of tracking points. At this point, the 3D points along a sequence that constructed from the images can be divided into two categories including static background and dynamic objects. In autonomous driving cases, only the object which moves quite differently from cameras and occupies larger area, regarded as main object, will mainly influence the estimation accuracy. While other moving objects can be rejected by common outlier rejection method such as [37]. We aim to cluster the sets of point cloud tracked along the sequence into two groups containing static camera motion and moving main object motion, respectively. That means, it is assumed to be two subspaces in the scene.

Although SSC is one of the best ways of motion segmentation, it is still a difficult task to work well in the high speed environment. In this section, we review the geometry of motion segmentation problem and show that how the 3D SSC method can address this problem.

Given a set of points $\left\{x_{ij} \in \mathbb{R}^2\right\}_{i=1,\ldots,P}^{j=1,\ldots,F}$ projected by 3D points $\left\{X_i \in \mathbb{R}^3\right\}_{i=1,\ldots,P}$ along moving coordinate frames $\left\{f_j\right\}_{j=1}^{F}$, all the feature points satisfy the affine projection model under a rigid-body motion:

$$\begin{bmatrix} x_{11} & \cdots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{F1} & \cdots & x_{FP} \end{bmatrix} = \begin{bmatrix} A_1 \\ \vdots \\ A_F \end{bmatrix} \begin{bmatrix} X_1 & \cdots & X_P \end{bmatrix} \tag{9}$$

where

$$A_j = K_j T_j = \begin{bmatrix} f/_{dX} & 0 & c_x \\ 0 & f/_{dX} & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} R_j & t_j \end{bmatrix} \in \mathbb{R}^{2\times4} \tag{10}$$

is the affine camera matrix at frame j. Since the 3D trajectories are acquired in Section III-B, there is no need to consider projection model. Therefore, the transformation matrix in 3D space:

$$T_j = \begin{bmatrix} R_j & t_j \end{bmatrix} \in \mathbb{R}^{3\times4} \tag{11}$$

can represent the transformation model of 3D points. Unlike affine subspaces segmentation, there is no risk that the segmentation of linear subspaces grouped by 3D transformation will incorrectly distinguish the subspaces because of affine structure of the data. With multiple motions in a scene, let $\{S_l\}_{l=1,\ldots,n}$ be a group of n linear subspace of $\mathbb{R}^{3\times F}$. Denote the matrix containing all the data points as

$$X \triangleq \begin{bmatrix} x_1 & \cdots & x_P \end{bmatrix} = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix} C \tag{12}$$

where $X_l \in \mathbb{R}^{3F\times N_l}$ is a data matrix of the points that lies in S_l and $C \in \mathbb{R}^{P\times P}$ is an unknown permutation matrix.

The SSC problem is the task of clustering the data encoded by sparse optimization program through spectral clustering framework. Vidal *et al.* [28] solved the optimization problem as

$$\min \|C\|_1 \quad s.t.\; X = XC,\, diag(C) = 0 \tag{13}$$

Then SSC method in [29] builds a graph G with the similarity matrix as $W = |C| + |C|^T$, which is used as the input of spectral clustering to infer the segmentation of data. It is proven that the SSC method is only suitable for low speed scenes. As the operation of taking absolute value for sparse representation matrix C, the objects that move in the opposite direction with the same speed as the camera or move along the camera's direction with twice speed will eventually get categorized as background because the negative coefficients disappeared. These kinds of situations often happens on highway or urban roads, where SSC method fails to detect moving objects. Hence, in this paper, we form a C matrix that every element in the matrix is set to be non-negative to improve the robustness in the case of high speed. Non-negative constraint was first proposed in [38] for sparse coding and then gets frequently used [39], [40]. It turns out that the non-negative constraint can learn the structure of data points effectively and induce a good result for classification. This constraint ensures that coefficients of the representation can be directly converted to graph weights, which also ensures that every data point is in the convex hull of its neighbours. Besides, the data points are prone to noise due to the errors in the processes of points tracking and depth estimation. Therefore, we use the Lasso optimization algorithm to recover the sparse solution with corrupted data points. The final modified optimization problem is expressed as:

$$\min\{\|C\|_1+\lambda\|X-XC\|_2\} \quad s.t.\; diag(C)=0,\, c_{ij}>0 \tag{14}$$

where the regularization parameter $\lambda > 0$ is a constant and the l_2 norm promotes having small entries in error $|X - XC|$. Note that this minimization problem is a convex problem, which can be solved using convex solver toolbox in Matlab as CVX. After solving the improved optimization program, we obtain a sparse representation for each point. Then a weighted graph G with similarity matrix $W \in \mathbb{R}^{P\times P}$ representing the weights of the edges is built to infer the segmentation by applying spectral clustering. The similarity matrix W is formed of normalized sparse coefficients C by setting $W = C + C^T$.

As the outliers are detected accurately, the rotation is refined and the translation is computed through minimizing reprojection error.

IV. EXPERIMENT RESULTS

In this section, the experiments are designed and carried out on the KITTI dataset [35] to evaluate our algorithm. Additionally, the improved SSC method is evaluated in the dynamic scene in KITTI dataset as well. We conduct all experiments on a computer with an Intel Core processor and 4 GB of memory in MATLAB implementations.

A. 3D SSC EVALUATION

For sequences in KITTI dataset, the feature trajectories are constructed using the probabilistic correspondences and ego-motion estimation. Visual examples of the 3D SSC method and the SSC method are given in Fig. 3. As can be observed in the first column, a significant number of points belong to the dynamic parts are detected by the 3D SSC method. The second column shows segmentation results obtained from the SSC method. The images in first row contain two cars in highway, while the scenes between second and fourth rows are in urban area where the vehicle moves at medium speed. The fifth row shows that the car drives in rural area and observes two bicycles which move much slower than the car itself. The camera in sixth row is completely static at a red light. Note that the SSC method fails to recognize the moving objects when the camera is moving as well, where the camera moves much faster than the moving handheld camera in Hokpins dataset [8]. Although the SSC method is proven to separate the trajectories better in completely static environment in the sixth row. On the contrary, the 3D SSC method solves the segmentation problem well in most cases even on the highway. Whereas the vehicle is seldom sitting idle on the road, the 3D SSC method can much meet the demand in practical cases.

B. VO EVALUATION

After evaluating the motion segmentation, we conduct the evaluation of our method on KITTI benchmark as well. This dataset provides 11 sequences that captured by driving car

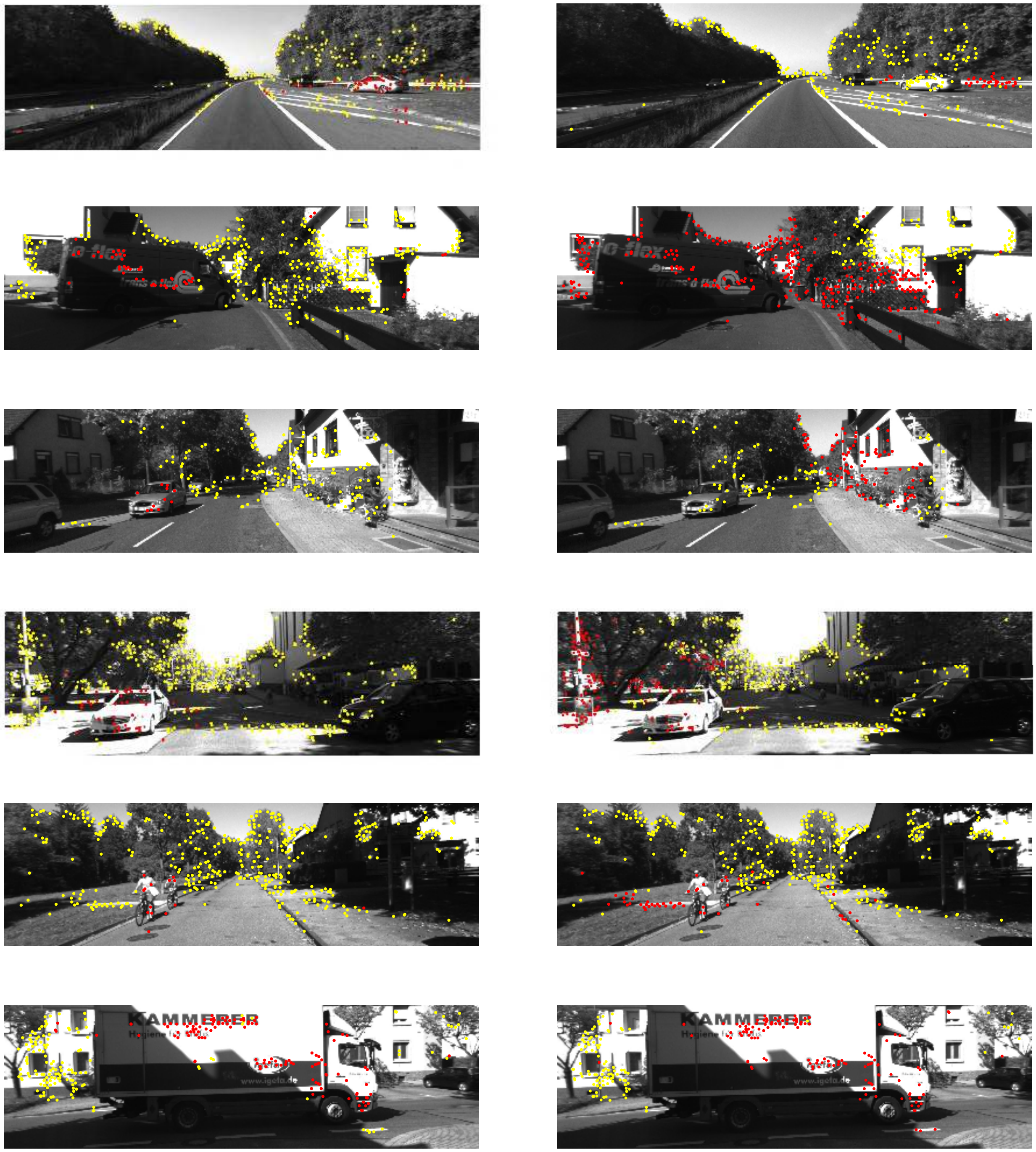


FIGURE 3. Visual comparison of 3D SSC and SSC. Sequences taken from KITTI. From left to right: 3D SSC method; SSC method. The first row is in highway; the second row is in low speed; the third and fourth row are in urban scenes with medium speed; while the fifth row is in rural scene and the dynamic objects moving much lower compared with the car and the camera is static in the last row.

around a city with ground truth for evaluation, in which the urban environments with high traffic is the main challenge for ego-motion estimation.

Several state-of-art VO and visual SLAM algorithms are compared in the subsequent experiments in order to evaluate the performance of our method. The slam algorithms contain: ORB-SLAM [41], the popular feature based visual odometry algorithm that achieves robust performance to large motion changes; SSLAM [42], [43], a visual odometry method that selects keyframes and tracks keypoints carefully and accurately to make the method more robust; ORB-SLAMM [44], a monocular SLAM system based on ORB-SLAM which can ensure mapping when the tracking fails. The visual odometry algorithms include: SDSO [45], a stereo visual odometry based on the known direct sparse odometry (DSO); PL-SVO [46], an extension work of the

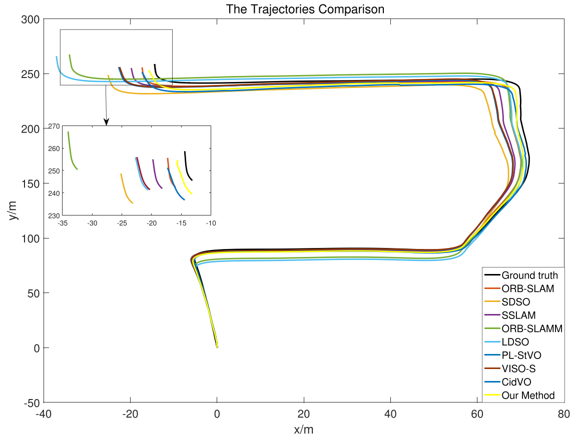


FIGURE 4. The reconstructed trajectories of the first sequence.

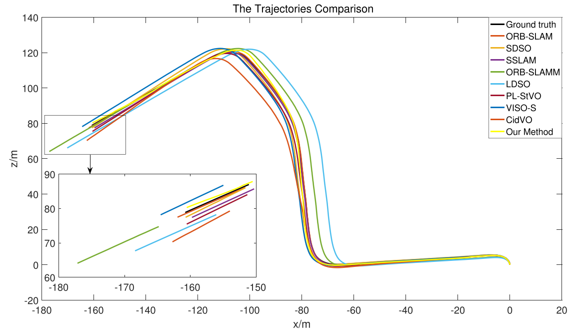


FIGURE 5. The reconstructed trajectories of the second sequence.

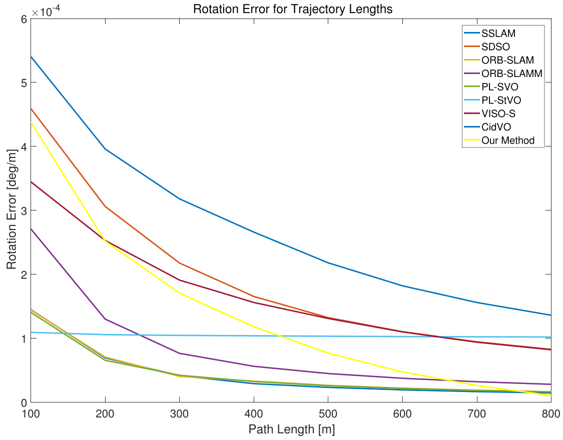


FIGURE 6. The rotation error for trajectory lengths.

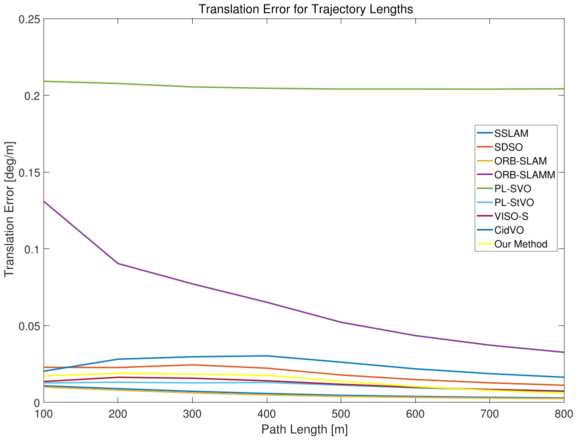


FIGURE 7. The translation error for trajectory lengths.

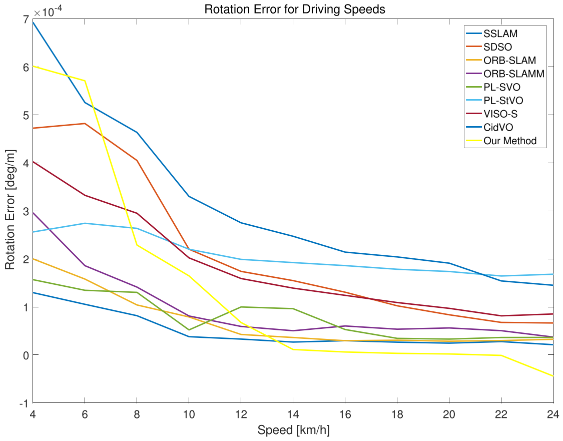


FIGURE 8. The rotation error for driving speeds.

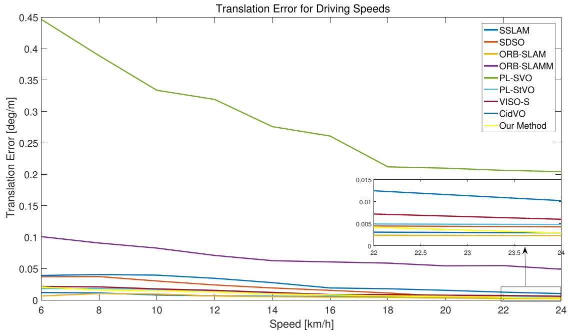


FIGURE 9. The translation error for driving speeds.

monocular semi-direct visual odometry (SVO) [47] which works with line segments; PL-StVO [48], a state-of-art stereo VO algorithm which achieves superior performance based on combination of both point and line segment features; VISO-S [35], one of the most popular stereo VO algorithms, which is usually chosen as a baseline method for comparison and evaluation; CidVO [37], the graph based stereo VO method utilizes a greedy algorithm to approximate the maximal clique in the graph to detect the set of inliers.

Fig. 4 and Fig. 5 show the trajectories reconstructed from our method compared with other methods. One can observe that our method performs best and possesses less cumulative error after accumulating both poses and errors over time. As the two sequences contain dynamic scenarios, the probabilistic and decoupled framework for rotation estimation and the 3D-SSC method for outliers removal both contribute to the performance.

For all sequences, the evaluation computes translation and rotation errors of length (100, 200, ..., 800) meters. Fig. 6-9 show the errors of experimental results of the sequence in Fig. 4. The errors at different trajectory lengths and driving speeds are plotted for each algorithm. On one hand,

TABLE 1. The statistical average errors of different methods in KITTI benchmark.

Methods	Rotation Error for lengths	Translation Error for lengths	Rotation Error for Speeds	Translation Error for Speeds
ORB-SLAM	0.000144	0.014265	0.000158	0.019709
SSLAM	0.000133	0.022988	0.000173	0.02235
ORB-SLAMM	0.000255	0.076931	0.000448	0.07089
PL-SVO	0.000148	0.142	0.000130	0.1140969
SDSO	0.00232	0.018605	0.000267	0.022441
PL-StVO	0.000572	0.023815	0.000347	0.024918
VISO-S	0.0002	0.02476	0.000247	0.026455
Cid-VO	0.000250	0.0242	0.000152	0.03296
Our Method	**0.0001705**	**0.01937**	**0.0002036**	**0.02147**

the rotation errors generated by our method decrease fastest along with the growth of both path length and speed. When the speed is high and the path length is long, the rotation errors of our method are smaller than other methods. On the other hand, the translation errors created by our method remain small among all algorithms, on account of the rejection of most outliers using 3D-SSC method.

The statistics errors according to the average of rotation and translation errors for all sequences in KITTI dataset (sequences 00-10) are reported in Table 1. Our method clearly outperforms all the visual odometry approaches in both rotation and translation errors. Besides, our method shows better results in translation against the slam algorithm SSLAM, which results from our 3D-SSC method for outliers rejection in dynamic environments. PL-SVO performs well in rotation because the line segments and direct visual odometry scheme provide robust tracking in high exposure scenarios, which are usual in KITTI or other outdoor scenes. But the lack of scale in this monocular system causes it performing worser by a large margin than other method. The typical feature based slam algorithm ORB-SLAM ranks first in both rotation and translation, which is based on robust ORB feature and pose optimization locally and globally. In general, our method is the top ranked in all visual odometry approaches and the second ranked of all vision methods.

V. CONCLUSION

In this paper, a novel algorithm for stereo visual odometry has been presented, which estimates the rotation ahead in probabilistic framework and detects dynamic objects based on 3D points especially when vehicle speed is fast. A more robust probabilistic framework for epipolar geometric estimation has been developed. In contrast to previous method, the proposed method has introduced the perpendicular distance to epipolar line and LMedS technique to achieve more robust estimation in the presence of noise. Given the essential matrix, the rotation extracted in advance is more accurate, without depth information participated in. Moreover, to remove disturbance of moving objects, the SSC algorithm has been modified with 3D points to adapt to fast movement of camera, which leads to a more accurate estimation of translation. The experimental results are evaluated with KITTI dataset. The results of 3D SSC show that the improved SSC method works well in dynamic environment. The proposed visual odometry algorithm performs better than the other algorithms. To improve the real-time performance of VO pipeline and study how to improve the robustness under high exposure conditions would be useful in the future.

REFERENCES

[1] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 55–81, 2015.

[2] T. Taketomi, H. Uchiyama, and S. Ikeda, "Visual SLAM algorithms: A survey from 2010 to 2016," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 3, p. 16, 2017.

[3] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, Dec. 2011.

[4] M. R. U. Saputra, A. Markham, and N. Trigoni, "Visual SLAM and structure from motion in dynamic environments: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, p. 37, 2018.

[5] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3354–3361.

[6] J. Janai, F. Güney, A. Behl, and A. Geiger. (2017). "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art." [Online]. Available: https://arxiv.org/abs/1704.05519

[7] A. Desai and D. J. Lee, "Visual odometry drift reduction using SYBA descriptor and feature transformation," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 3, pp. 1839–1851, Jul. 2016.

[8] *Results on the Hopkins 155 Dataset*. Accessed: Aug. 28, 2016. [Online]. Available: http://www.vision.jhu.edu/motion.php

[9] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2004, pp. 652–659.

[10] A. Pumarola, A. Vakhitov, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer, "PL-SLAM: Real-time monocular visual SLAM with points and lines," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Singapore, May 2017, pp. 4503–4508.

[11] S. L. von, V. Usenko, and D. Cremers. (2018). "Direct sparse visual-inertial odometry using dynamic marginalization." [Online]. Available: https://arxiv.org/abs/1804.05625

[12] M. A. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[13] J. Deigmoeller and J. Eggert, "Stereo visual odometry without temporal filtering," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2016.

[14] M. Persson, T. Piccini, M. Felsberg, and R. Mester, "Robust stereo visual odometry from monocular techniques," in *Proc. Intell. Vehicles Symp. (IV)*, 2015, pp. 686–691.

[15] M. Buczko and V. Willert, "Flow-decoupled normalized reprojection error for visual odometry," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 1161–1167.

[16] I. Cvišić and I. Petrović, "Stereo odometry based on careful feature selection and tracking," in *Proc. Eur. Conf. IEEE Mobile Robots (ECMR)*, Sep. 2015, pp. 1–6.

[17] B. Guan, P. Vasseur, C. Demonceaux, and F. Fraundorfer, "Visual odometry using a homography formulation with decoupled rotation and translation estimation using minimal solutions," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, 2018.

[18] J. Domke and Y. Aloimonos, "A probabilistic framework for correspondence and egomotion," in *Dynamical Vision*. Berlin, Germany: Springer, 2007, pp. 232–242.

[19] H. Shah and A. Lakshmikumar, "Probabilistic egomotion from a statistical framework," in *Proc. BMVC*, 2007, pp. 1–10.

[20] D.-J. Lee, P. Merrell, Z. Wei, and B. E. Nelson, "Two-frame structure from motion using optical flow probability distributions for unmanned air vehicle obstacle avoidance," *Mach. Vis. Appl.*, vol. 21, no. 3, pp. 229–240, 2010.

[21] D.-J. Lee, P. C. Merrell, B. E. Nelson, and W. Wei, "Multi-frame structure from motion using optical flow probability distributions," *Neurocomputing*, vol. 72, nos. 4–6, pp. 1032–1041, 2009.

[22] H. Silva, A. Bernardino, and E. Silva, "A voting method for stereo egomotion estimation," *Int. J. Adv. Robotic Syst.*, vol. 14, no. 3, p. 1729881417710795, 2017.

[23] D. Barnes, W. Maddern, G. Pascoe, I. Posner. (2017). "Driven to distraction: Self-supervised distractor learning for robust monocular visual odometry in urban environments." [Online]. Available: https://arxiv.org/abs/1711.06623

[24] H. Azartash, K.-R. Lee, and T. Q. Nguyen, "Visual odometry for RGB-D cameras for dynamic scenes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1280–1284.

[25] D.-H. Kim, S.-B. Han, and J.-H. Kim, "Visual odometry algorithm using an RGB-D sensor and IMU in a highly dynamic environment," in *Robot Intelligence Technology and Applications 3*. Cham, Switzerland: Springer, 2015, pp. 11–26.

[26] K. Lenac, I. Maurović, and I. Petrović, "Moving objects detection using a thermal camera and IMU on a vehicle," in *Proc. Int. Conf. IEEE Elect. Drives Power Electron. (EDPE)*, Sep. 2015, pp. 212–219.

[27] D.-H. Kim and J.-H. Kim, "Effective background model-based RGB-D dense visual odometry in a dynamic environment," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1565–1573, Dec. 2016.

[28] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2003.

[29] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.

[30] Y.-X. Wang and H. Xu, "Noisy sparse subspace clustering," *J. Mach. Learn. Res.*, vol. 17, no. 3, pp. 320–360, 2016.

[31] C. Yang, D. Robinson, and R. Vidal, "Sparse subspace clustering with missing entries," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2463–2472.

[32] C. You, D. Robinson, and R. Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3918–3927.

[33] C.-G. Li and R. Vidal, "Structured sparse subspace clustering: A unified optimization framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 277–286.

[34] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 430–443.

[35] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3D reconstruction in real-time," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 963–968.

[36] M. Rais, G. Facciolo, E. Meinhardt-Llopis, J.-M. Morel, A. Buades, B. Coll. (2017). "Accurate otion estimation through random sample aggregated consensus." [Online]. Available: https://arxiv.org/abs/1701.05268

[37] A. Howard, "Real-time stereo visual odometry for autonomous ground vehicles," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2008, pp. 3946–3952.

[38] P. O. Hoyer, "Modeling receptive fields with non-negative sparse coding," *Comput. Neurosci., Trends Res.*, vols. 52–54, pp. 547–552, Jun. 2003.

[39] L. Zhuang, H. Gao, Z. Lin, Y. Ma, X. Zhang, and N. Yu, "Non-negative low rank and sparse graph for semi-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2328–2335.

[40] R. He, W.-S. Zheng, B.-G. Hu, and X.-W. Kong, "Nonnegative sparse coding for discriminative semi-supervised learning," in *Proc. CVPR*, 2011, pp. 792–801.

[41] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, Oct. 2017.

[42] M. Fanfani, F. Bellavia, and C. Colombo, "Accurate keyframe selection and keypoint tracking for robust visual odometry," *Mach. Vis. Appl.*, vol. 27, no. 6, pp. 833–844, 2016.

[43] F. Bellavia, M. Fanfani, and C. Colombo, "Selective visual odometry for accurate AUV localization," *Auto. Robots*, vol. 41, no. 3, pp. 133–143, 2017.

[44] H. A. Daoud, A. Q. M. Sabri, C. K. Loo, and A. M. Mansoor, "SLAMM: Visual monocular SLAM with continuous mapping using multiple maps," *PLoS ONE*, vol. 13, no. 4, p. e0195878, 2018, doi: 10.1371/journal.pone.0195878.

[45] J. Engel, V. Koltun, and D. Cremers. (2016). "Direct sparse odometry." [Online]. Available: https://arxiv.org/abs/1607.02565

[46] R. Gomez-Ojeda, J. Briales, and J. Gonzalez-Jimenez, "PL-SVO: Semi-direct monocular visual odometry by combining points and line segments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2016, pp. 4211–4216.

[47] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2014, pp. 15–22.

[48] R. Gomez-Ojeda and J. Gonzalez-Jimenez, "Robust stereo visual odometry through a probabilistic combination of points and line segments," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, May 2016, pp. 2521–2526.

YAN WANG received the Ph.D. degree in control theory and control engineering from Northeastern University, Shenyang, China, in 2003. From 2003 to 2005, she was a Postdoctoral Fellow with the National Key Laboratory of Intelligent Technology and Systems, Tsinghua University. She is currently an Associate Professor with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. Her research interests include networked control systems, intelligent systems, and computation.

HUI-QI MIAO received the bachelor's degree in automation from the Beijing University of Chemical Technology, Beijing, China, in 2016. She is currently pursuing the M.S. degree with the Intelligent Control and Computation Laboratory, Beihang University, Beijing. Her research interests include pattern recognition and machine vision.

LEI GUO received the B.S. and M.S. degrees from Qufu Normal University, China, in 1988 and 1991, respectively, and the Ph.D. degree in control engineering from Southeast University, in 1997. Since 2007, he has been a "Lantian" Distinguished Professor with Beihang University. He has published more than 100 referred papers and one monograph with Springer. His research interests include robust control and filtering, stochastic systems, fault detection, and nonlinear control with their applications to aerospace systems. He served as an editorial board member for four journals and three conferences.