Three-Point Direct Stereo Visual Odometry

Jeong-Kyun Lee
leejk@gist.ac.kr
Kuk-Jin Yoon
kjyoon@gist.ac.kr

Computer Vision Laboratory
Gwangju Institute of Science and
Technology (GIST), South Korea

Abstract

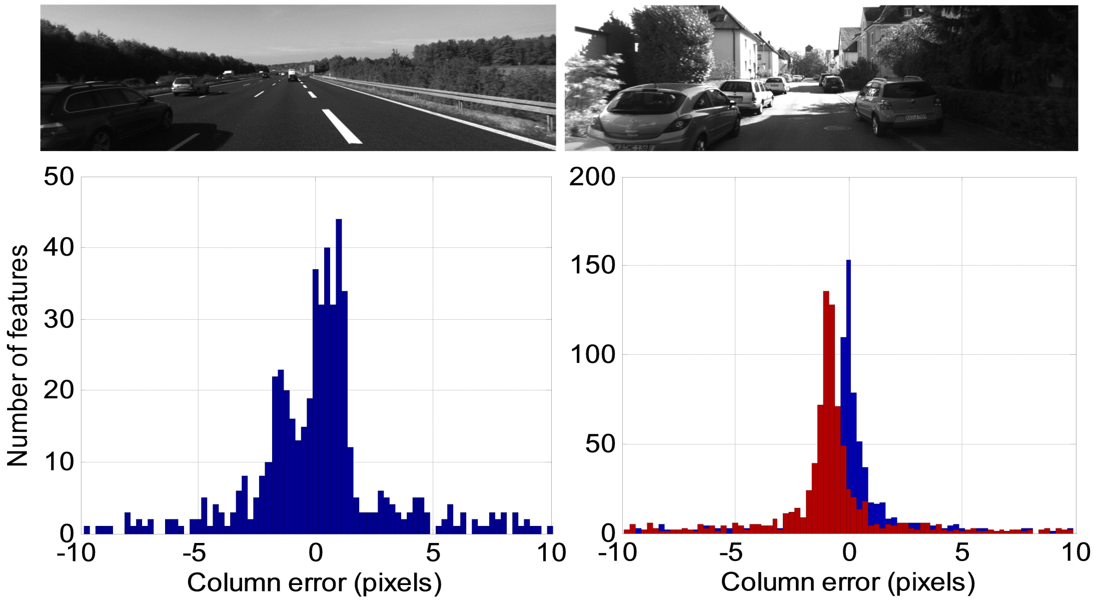
Stereo visual odometry estimates the ego-motion of a stereo camera given an image sequence. Previous methods generally estimate the ego-motion using a set of inlier features while filtering out outlier features. However, since the perfect classification of inlier and outlier features is practically impossible, the motion estimate is often contaminated by erroneous inliers. In this paper, we propose a novel three-point direct method for stereo visual odometry, which is more accurate and robust to outliers. To improve both accuracy and robustness, we consider two key points: sampling a minimum number of features, *i.e.*, 3 points, and minimizing photometric errors in order to maximally reduce measurement errors. In addition, we utilize temporal information of features, *i.e.*, feature tracks. Local features are updated by the feature tracks and the updated feature points improve the performance of the proposed pose estimation. We compare the proposed method with other state-of-the-art methods and demonstrate the superiority of the proposed method through experiments on the KITTI benchmark.

1 Introduction

Estimating the ego-motion of a camera given an image sequence, called visual odometry, is one of the main research topics in computer vision and robotics. It is extensively used in a large number of applications such as object detection and tracking [3] and navigation [17]. Especially, visual odometry based on a stereo camera has been intensively studied because it is able to accurately and instantly estimate the ego-motion at a metric scale from calibrated stereo configuration and so is considered as a powerful means for autonomous driving.

Existing visual odometry methods are generally classified into two groups; feature-based and direct methods. Feature-based methods [1, 2, 5, 8, 13, 21, 24, 26, 27] exploit feature points extracted and tracked in consecutive frames. The measurement errors are defined by the re-projection errors of feature points, and the ego-motion is estimated by minimizing the re-projection errors of feature points with the assumption that the measurement errors of tracked feature points usually conform to the Laplace distribution [2]. However, since the local appearance of the feature point consecutively varies with time, the location of the tracked feature point tends to drift from its initial location during tracking and this results in error accumulation. In contrast, direct methods [7, 9] minimize the measurement errors measured by intensity differences between consecutive images. Since a large number of pixels are utilized for motion estimation, these methods are superior to feature-based methods in the error accumulation aspect. However, in return, the direct methods have a difficulty

Pages 83.1-83.12
DOI: https://dx.doi.org/10.5244/C.30.83



(a) (b)

Figure 1: The distributions of re-projection errors for estimated (blue) and ground truth motion (red) in selected frames of the KITTI dataset. In (a), the distribution violates the Laplace distribution assumption of the measurement errors because of outliers from moving objects. In (b), the measurement errors are biased.

in handling outliers and are vulnerable to illumination change. For this reason, most of the recent visual odometry methods are feature-based and exploit local image features for robustness in real-world scenes. They generally use a maximum consensus set of inlier feature points on the Laplace distribution assumption of measurement errors. However, as shown in Fig. 1, the assumption is often violated in the real-world scene because of lots of outliers and/or biases of measurement errors. In this situation, although a motion candidate may be accurately obtained from a small number of point samples in the RANSAC framework [10], its final estimate optimized from the set of inliers can be rather erroneous. The work of Chum *et al.* [6] tried to solve the similar problem using an iterative local RANSAC scheme but it is still hard to hold thoroughly uncontaminated inliers only.

In this paper, we propose a new hybrid method of the feature-based and direct methods, which samples and uses a minmimum number of feature points only (*i.e.*, 3 points) to estimate ego-motion without any additional optimization using all inliers. The proposed method maximally excludes the errors caused by inaccurate inliers while handling the sensitivity to measurement errors caused by using a small number of feature points. Experimental results show the proposed method produces better performance than the existing methods using all inliers.

1.1 Related work

Since the work of Nister *et al.* [24], visual odometry has been intensively studied in the last decade. Because we focus on stereo visual odometry in this paper, we review the works on stereo visual odometry. More details on visual odometry can be found in [12, 30].

As mentioned earlier, methods for stereo visual odometry are usually divided into two classes; feature-based and direct methods. The basis of the feature-based methods has been formed by Nister *et al.* [24]. In the work, local features extracted by the Harris corner detector [16] are matched using normalized cross-correlation (NCC) and ego-motion is estimated using the perspective-3-point (P3P) algorithm [15] and the RANSAC technique [10].

Geiger *et al*. [13] proposed an algorithm that uses a simplified feature detection and matching method to quickly handle many features and performs the Gauss-Newton optimization algorithm to minimize the re-projection errors of feature points extracted in both stereo image pairs. In [5, 29], a key-frame scheme used in the monocular visual odometry was exploited. It is computationally efficient and possible to estimate more accurate motion from the feature points with sufficient displacement. In recent works [1, 2, 21, 26], the trajectories of features tracked over several frames are taken into account to update the 3D positions of the features. For the feature update, the extended Kalman filter (EKF) is used in [1] and the local bundle adjustment (LBA) [32] is used in [21, 26]. They significantly improve the accuracy of visual odometry holding high ranks in the visual odometry benchmark [14].

On the other hand, the direct methods [7, 9, 11] have attracted attention in recent years because of the advantages in both computational efficiency and accuracy aspects. Comport *et al*. [7] proposed the direct stereo visual odometry based on the quadrifocal tensor. The algorithm achieves the highest accuracy among the algorithms that do not utilize temporal feature information. Engel *et al*. [9] proposed a semi-dense direct method that updates features at each frame and thus improves accuracy significantly. However, the previous direct methods still have a difficulty in determining outliers because the M-estimator in [7] or the simple outlier rejection method in [9] are not robust for the scenes with many outliers. In the next section, we propose a novel hybrid method that takes advantages of the feature-based and direct methods.

2 Proposed method

To solve the problem shown in Fig. 1, we propose a method aiming at estimating ego-motion using only uncontaminated feature points. As it is difficult to distinguish the uncontaminated features among lots of features, we randomly sample a minimum number of features and exploit them for the estimation. However, using fewer features in feature-based methods can rather degrade the performance because measurement errors caused in feature matching and tracking can be propagated to the ego-motion estimation directly. Therefore, we propose a hybrid method of feature-based and direct methods. It minimizes measurement errors by the ego-motion estimation directly using image intensity values without any intermediate step to search measurements and is more robust than conventional direct methods.

2.1 Feature extraction

To initialize features, key points are extracted using the FAST corner detector [28]. The key points with local maximum responses within an 11×11 window are selected by a non-maximum suppression algorithm [23]. The points extracted from stereo image are matched using image descriptors. The descriptor is a 7×7 patch centered at each key point in a gradient image computed by the 3×3 Sobel operator. The descriptors are matched by the sum of absolute differences (SAD). We then compute disparities of key points with sub-pixel accuracy. The key points with the reliable disparities are accepted as features and we keep about 1000 features per frame.

2.2 Ego-motion estimation

2.2.1 Modeling

Given the pixel locations and 3D positions of feature points in the previous image, the direct method estimates the ego-motion that minimizes photometric errors between the image

Algorithm 1 Three-point direct method

```
Input: $F, I_{k-1}, I_k$
Output: $\xi_{k,k-1}$
$n_{max} \leftarrow 0$
repeat
    $\xi \leftarrow \mathbf{0}_{6\times 1}$
    sample 3 feature points $\bar{F} \subset F$
    for $l = MaxPyramidLevel$ to 1 do
        determine $\mathcal{R}_{f_i}$ for $\forall f_i \in \bar{F}$
        repeat
            compute $J_{\mathbf{p}}$, $H$, and $I_{k-1}(\mathbf{p}) - I_k(\pi(\mathbf{X_p}, \mathbf{T}(\xi)))$ for $\forall \mathbf{p} \in \mathcal{R}_{f_i}$
            compute $\Delta\xi$ using Eq. (4)
            update $\xi$ with $\Delta\xi$ using Eq. (2)
        until $N$ times or $\xi$ converges
    end for
    $n \leftarrow$ NumberOfInliers($\xi$)
    if $n_{max} < n$ then
        $n_{max} \leftarrow n$, $\xi_{best} \leftarrow \xi$
    end if
until $M$ times
$\xi_{k,k-1} \leftarrow \xi_{best}$
```

patches of the features. This problem is concerned with finding a least-squares estimate of ego-motion ξ. The cost function to be minimized is defined as

$$C(\xi) = \frac{1}{2}\sum_{i}^{n}\sum_{\mathbf{p}}^{\mathcal{R}_i}\{I_k(\pi(\mathbf{X_p}, \mathbf{T}(\xi))) - I_{k-1}(\mathbf{p})\}^2, \tag{1}$$

where $\xi = [\mathbf{v}^\top, \omega^\top]^\top \in \mathfrak{se}(3)$ is a 6D pose comprised of a linear velocity $\mathbf{v} \in \mathbb{R}^3$ and an angular velocity $\omega \in \mathbb{R}^3$, n is the number of feature points, $\mathcal{R}_i$ is a set of pixels around the i-th feature point (*i.e.*, an image patch of the i-th feature), $\mathbf{p} \in \Omega$ is a 2D coordinate of a pixel in the image domain, $I_k(\mathbf{p}) : \Omega \to \mathbb{R}$ is an intensity of a pixel $\mathbf{p}$ in the k-th frame, $\mathbf{T}(\xi) : \mathfrak{se}(3) \to SE(3)$ is the exponential map, and $\pi(\mathbf{X_p}, \mathbf{T}(\xi)) : \mathbb{R}^3 \times SE(3) \to \Omega$ is a projection function. The 3D cartesian coordinate $\mathbf{X_p}$ is transformed by $\mathbf{T}(\xi)$ and then projected onto the image coordinate $\mathbf{p}$.

To minimize Eq. (1), we use the Gauss-Newton method. The pose ξ is iteratively updated with the increment of the parameter $\Delta\xi$. The parameter update is performed by the forward compositional method [4, 11] as in Eq. (2).

$$\mathbf{T}(\xi) \longleftarrow \mathbf{T}(\xi)\mathbf{T}(\Delta\xi) \tag{2}$$

If Eq. (2) is substituted in Eq. (1), then we obtain a function of the increment $\Delta\xi$. Since the altered cost function is nonlinear, we approximate the value I_k with its first-order Taylor expansion as

$$C(\Delta\xi) \approx \sum_{i}^{n}\sum_{\mathbf{p}}^{\mathcal{R}_i}\{I_k(\pi(\mathbf{X_p}, \mathbf{T}(\xi))) + J_{\mathbf{p}}\Delta\xi - I_{k-1}(\mathbf{p})\}^2, \text{ where } J_{\mathbf{p}} = \left.\frac{\partial I_k(\mathbf{p})}{\partial \Delta\xi}\right|_{\Delta\xi=\mathbf{0}}. \tag{3}$$

The jacobian $J_{\mathbf{p}}$ is consequently linear and the increment $\Delta\xi$ is therefore computed as

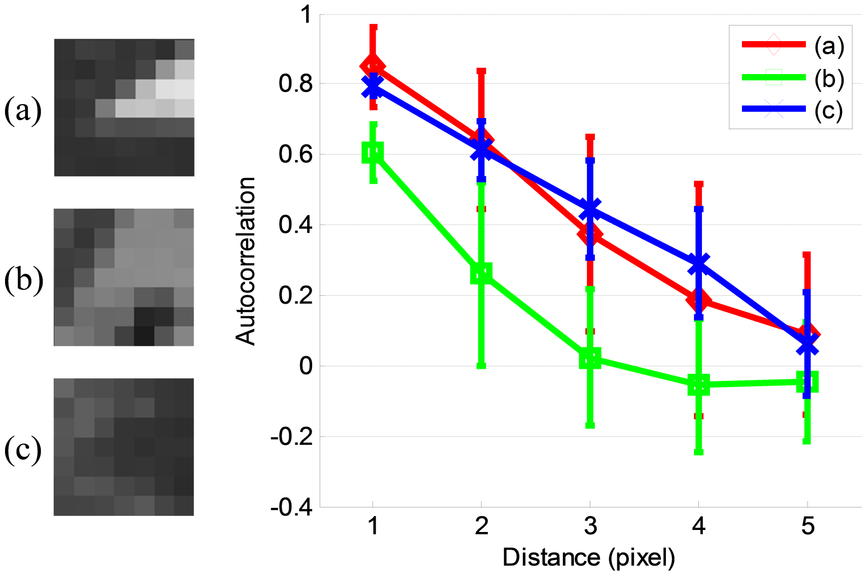


Figure 2: The means and standard deviations of autocorrelation values for the different image patches. The autocorrelation of an image patch is commonly decreased along the distance and the image patches with different intensity patterns have different auotocorrelation values along the distance. Thus, the autocorrelation values can be used as an adaptive threshold to match image patchs.

$$\Delta\xi = H^{-1}\sum_{i}^{n}\sum_{\mathbf{p}}^{\mathcal{R}_i} J_{\mathbf{p}}^{\mathrm{T}}\left[I_{k-1}(\mathbf{p}) - I_k(\pi(\mathbf{X}_{\mathbf{p}},\mathbf{T}(\xi)))\right], \text{ where } H = \sum_{i}^{n}\sum_{\mathbf{p}}^{\mathcal{R}_i} J_{\mathbf{p}}^{\mathrm{T}} J_{\mathbf{p}}. \tag{4}$$

Since the direct methods are usually suitable for small displacement of feature points but not for large displacement, we exploit a hierarchical process using image pyramids to handle large motion. In addition, we implement the proposed method with an inverse compositional method. It is computationally more efficient than the forward compositional method as demonstrated in [4]. The details are referred to [4, 11].

2.2.2 Three-point direct method

To accurately estimate the camera pose using a small number of feature points, we propose the hybrid method, *i.e.*, 3-point direct stereo visual odometry. The proposed method is shown in Algorithm 1. Given a set of features $F = \{f_1, ..., f_n\}$, we randomly sample three features. Then, the direct method described in Sec. 2.2.1 is performed using the three sampled features. In the experiment, we set *MaxPyramidLevel* to 5, which is enough to handle large displacement of features. The number of iterations N is set to 30 but the iteration is terminated if the parameter ξ converges. The sizes of the image patches for $\mathcal{R}_{f_i}$ are 9×9, 9×9, 7×7, 7×7, and 7×7 windows in the ascending order of the pyramid levels, considering the tradeoff between computational speed and accuracy. Assuming each image patch to be fronto-parallel, all the disparities in an image patch are also assumed to be the same.

At line 13 of Algorithm 1, whether a feature is an inlier or not is determined by matching the image patches of the feature point projected onto I_{k-1} and I_k, given the motion candidate ξ. If the zero-mean normalized cross-correlation (ZNCC) between the 9×9 image patches for the i-th feature is larger than a threshold T_i, then the i-th feature is accepted as an inlier. We make the threshold T_i adaptively vary because a constant threshold is not proper for checking the correlation between the patches with different shapes and intensity values. Accordingly, we adopt the self-aware distance transform [25] that adaptively measures a matching threshold from autocorrelation of each image patch. As shown in Fig. 2, the patches with different intensity patterns have different autocorrelation values. If we assume that an image patch changes smoothly over an image sequence, the autocorrelation of the patch can be used to predict the matching threshold. When μ_i and σ_i denote the mean and standard deviation of the autocorrelation values of the i-th feature for the distance of 1 pixel, the threshold T_i is defined as in Eq. (5).

$$T_i = \begin{cases} \alpha, & \text{if } \mu_i + 2\sigma_i > \alpha \\ \beta, & \text{if } \mu_i + 2\sigma_i < \beta \\ \mu_i + 2\sigma_i, & \text{otherwise} \end{cases} \tag{5}$$

The parameters α and β are upper and lower bound thresholds empirically set to 0.97 and 0.7, respectively. The adaptive thresholding scheme improves the accuracy and robustness of visual odometry, which will be demonstrated in Sec. 3.

2.3 Feature update

The 3D positions of features are computed by triangulation. However, as mentioned in the literature [5, 18], the position of a feature computed from stereo matching is not accurate because of inaccurate stereo calibration and the uncertainty of the depth at a far distance. Therefore, it is necessary to correct 3D positions of the features. In order to update the feature, we exploit the inverse depth parameterization [22]. The parameter of the i-th feature is defined as $\mathbf{y}_i = [X_i, Y_i, Z_i, \theta_i, \phi_i, \rho_i]^\top$ where the vector $[X_i, Y_i, Z_i]^\top$ is the position of the camera, θ_i and ϕ_i azimuth and elevation, and ρ_i an inverse depth. From the parameter $\mathbf{y}_i$, the 3D location of the i-th feature is computed as follows:

$$\mathbf{X}_i = [X_i, Y_i, Z_i]^\top + \frac{1}{\rho_i}\mathbf{m}(\theta_i, \phi_i), \text{ where } \mathbf{m} = [\cos\phi_i \sin\theta_i, -\sin\phi_i \cos\theta_i, \cos\theta_i]^\top. \tag{6}$$

The parameter $\mathbf{y}_i$ is updated by the EKF. A measurement equation of the EKF is defined by the projection equation $\mathbf{h}_i = \pi(\mathbf{X}_i, \mathbf{T}_{cw})$ with the Euclidean transformation $\mathbf{T}_{cw}$ from the world coordinates to the camera coordinates. Measurements are obtained from the Kanade-Lucas-Tomasi (KLT) feature tracker [31]. If a measurement of a feature is not acquired, then the feature is removed from a list of features.

2.4 Implementation details

Illumination correction In practical situations, illumination of the scene can be varied over time. It may significantly affect the proposed method because both the direct visual odometry and the KLT tracker are based on the intensity conservation assumption. Hence, to correct the illumination change, we properly exploit the global color transfer scheme [20]. This scheme assumes that global illumination change between two images conforms to affine transformation. The affine transformation is defined as

$$I_{k-1} = \frac{\sigma_{k-1}}{\sigma_k}(I_k - \mu_k) + \mu_{k-1}, \tag{7}$$

where μ_k and σ_k are the mean and the standard deviation of all the intensity values of an image at time k. Although the illumination is not perfectly corrected between two consecutive images, the distinctiveness of corner features is maintained enough under the approximately matched illumination. We observe that the proposed method successfully works in various test sets after the affine global illumination transformation model is applied.

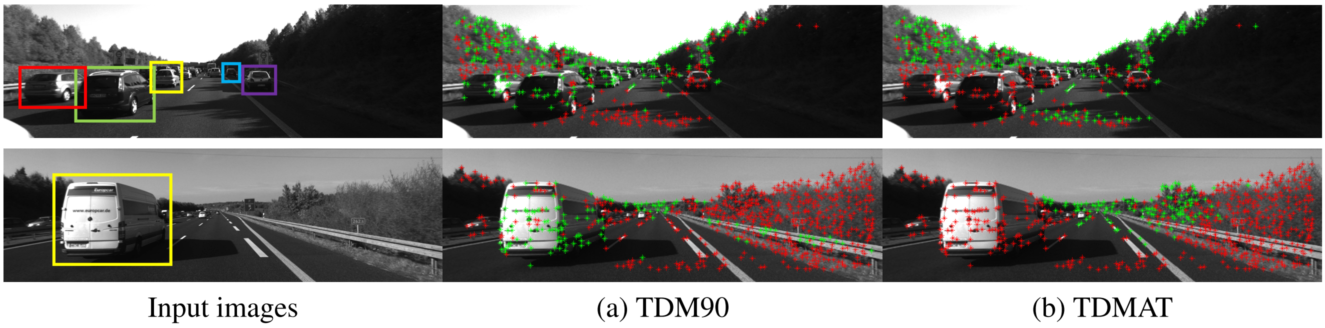


Figure 3: The robustness of the adaptive thresholding scheme. In (a) and (b), green and red points represent inliers and outliers, respectively. The cars in the colored boxes are moving objects on which the feature points should be classified into outliers.

Motion prediction Motion prediction can be used to guide the ego-motion estimation. We assume that motion is changed with a constant linear and angular velocity. For the sake of the simplicity of the implementation, the motion estimated from the previous two frames is used as the prediction of the current ego-motion. The motion prediction enables the ego-motion to be successfully estimated even in repetitive patterns and reduces computational time.

Feature selection It is important to determine which features should be sampled in the RANSAC process of the ego-motion estimation, because we suppose that motion can be reliably computed using a small number of features. Cvisic and Petrovic [8] exploited a survival age of a feature for careful feature selection which remarkably improves the accuracy of visual odometry. We also consider the survival age but its use is slightly different from the method of [8]. In the RANSAC process, we use only the features over the age of 1 in order to sample 3 points and check inlier features. Since most outlier features are immediately removed, survival features can be considered reliable. It increases the probability of selecting uncontaminated features.

3 Experimental results

We evaluate the proposed method using the publicly available KITTI dataset [14]. The benchmark datasets include a lot of challenging image sequences such as country roads and highways in which the illumination is considerably varied or there exist only a small number of reliable inlier features because of moving objects. The proposed method is compared with the VISO2-S [13] and the baseline algorithm. The baseline algorithm uses the feature extraction and matching methods mentioned in Sec. 2.1. In the algorithm, a maximum consensus set of inlier features is selected in the RANSAC process using the robust P3P algorithm [19] and the motion estimate is then optimized by the Gauss-Newton method using all the inliers. Base1 and Base3 denote the baseline algorithm with inlier thresholds of 1 and 3 pixels, respectively. TDM70 and TDM90 denote the three-point direct method with ZNCC thresholds of 0.70 and 0.90, respectively. TDMAT is the three-point direct method with the adaptive thresholding scheme. All the RANSAC processes of both the baseline and proposed algorithms equally run 1000 iterations at each frame.

3.1 Performance analysis

Table 1 shows the performances of the baseline and the proposed algorithms. The table represents an average of relative position errors, which is the main ranking criterion in the

Table 1: Performance comparison in the KITTI dataset. MP denotes the motion prediction, U the feature update, and S the feature selection technique. In this table, we only represent averages of relative translation errors (%).

Algorithm	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
VISO2-S [13]	2.74	4.65	2.28	2.48	0.99	2.14	1.30	2.29	2.78	2.55	1.62	2.53
Base1	2.21	6.16	1.65	0.91	1.10	1.37	1.35	2.24	1.60	1.53	0.86	1.90
Base3	2.44	13.06	1.44	1.22	1.29	1.43	1.37	2.73	2.10	1.76	2.15	2.40
TDM70	2.00	10.97	1.64	1.38	0.98	1.42	1.22	1.68	1.88	1.53	0.82	2.12
TDM90	1.64	8.44	1.55	1.07	1.26	1.24	1.48	1.47	1.47	1.68	0.66	1.81
TDMAT	1.74	4.87	1.57	1.07	1.01	1.25	1.45	1.68	1.52	1.54	0.51	**1.67**
Base1+MP	2.26	4.55	1.65	0.93	1.10	1.42	1.36	2.27	1.63	1.54	0.84	1.85
TDMAT+MP	1.76	3.60	1.52	1.14	0.95	1.20	1.34	1.80	1.50	1.51	0.57	**1.59**
Base1+MP+U	1.44	4.65	1.15	0.91	1.10	0.92	1.11	1.56	1.28	1.40	0.66	1.38
TDMAT+MP+U	1.33	4.12	1.10	1.07	1.11	0.88	0.96	1.41	1.22	1.39	0.46	**1.29**
Base1+MP+U+S	1.07	4.61	0.88	1.05	1.31	0.84	1.11	1.23	1.18	1.38	0.67	1.20
TDMAT+MP+U+S	1.04	4.05	0.99	1.23	1.53	0.87	1.05	1.14	1.13	1.42	0.60	**1.18**

KITTI benchmark [14]. In addition, Fig. 4 shows some results of the proposed method for the KITTI benchmark dataset.

Adaptive thresholding scheme We qualitatively evaluate the adaptive thresholding scheme in some challenging image sequences. In Fig. 3, there exist moving objects on which there are many feature points expected to be outliers. Contrary to the results of TDM90, TDMAT correctly classifies all the points on the moving cars into the outliers. These results demonstrate that the adaptive thresholding scheme improves the classification performance. In addition, the performance of the proposed method is quantitatively evaluated in Table 1. TDMAT shows the lowest translation error of 1.67% among VISO2-S, the baseline, and the proposed algorithm. It demonstrates that the proposed 3-point direct method achieves better performance than the algorithms using all inliers and the adaptive thresholding scheme is better than the constant thresholding scheme as well.

Superiority of the three-point direct method We have assumed that the ego-motion computed from a minimum number of good features will be more accurate than the estimate from a lot of moderate-quality inliers. It can be verified by comparing TDMAT and Base1. As shown in Table 1, TDMAT is superior to Base1 in most of the sequences. In particular, the proposed method shows the better performance in the urban scenes such as the dataset 0, 5, 7, and 10 because a lot of distinctive feature points extracted in the urban scenes improve the performance of the 3-point direct method.

Motion prediction In the sequence 1 of Table 1, the translation errors of all the methods are very high because feature matching fails due to the repetitive patterns such as guard rails. The motion prediction is useful in this case. It does not usually influence the performance but helps the ambiguous features matched. In Table 1, it is verified that the performances of both Base1+MP and TDMAT+MP are higher than those of Base1 and TDMAT.

Feature update and selection The feature update and selection schemes are evaluated in Table 1. The performances of TDMAT+MP+U and TDMAT+MP+U+S are remarkably improved in comparison with TDMAT+MP. Although the feature selection scheme considering

Table 2: KITTI leaderboard. The main techniques applied to each method are represented in the bracket. BA represents the local bundle adjustments, L the loop closing, MP the motion prediction, U the feature update using feature tracks, S the feature selection, and K the key-frame scheme.

Method	Translation	Rotation
SOFT (S) [8]	1.03 %	0.0029 [deg/m]
cv4xv1-sc (BA,MP) [26]	1.09 %	0.0029 [deg/m]
S-LSD-SLAM (L,U,K) [9]	1.20 %	0.0033 [deg/m]
MFI (U) [2]	1.30 %	0.0030 [deg/m]
S-PTAM (BA,MP,K) [27]	1.35 %	0.0023 [deg/m]
TLBBA (BA)[21]	1.36 %	0.0038 [deg/m]
Proposed (MP, U, S)	1.47 %	0.0030 [deg/m]
StereoSFM (U)[1]	1.51 %	0.0042 [deg/m]
SSLAM (S,K) [5]	1.57 %	0.0044 [deg/m]
eVO (MP,K)[29]	1.76 %	0.0036 [deg/m]
Proposed (without MP, U, S)	1.97 %	0.0036 [deg/m]
D6DVO [7]	2.04 %	0.0051 [deg/m]
VISO2-S [13]	2.44 %	0.0114 [deg/m]

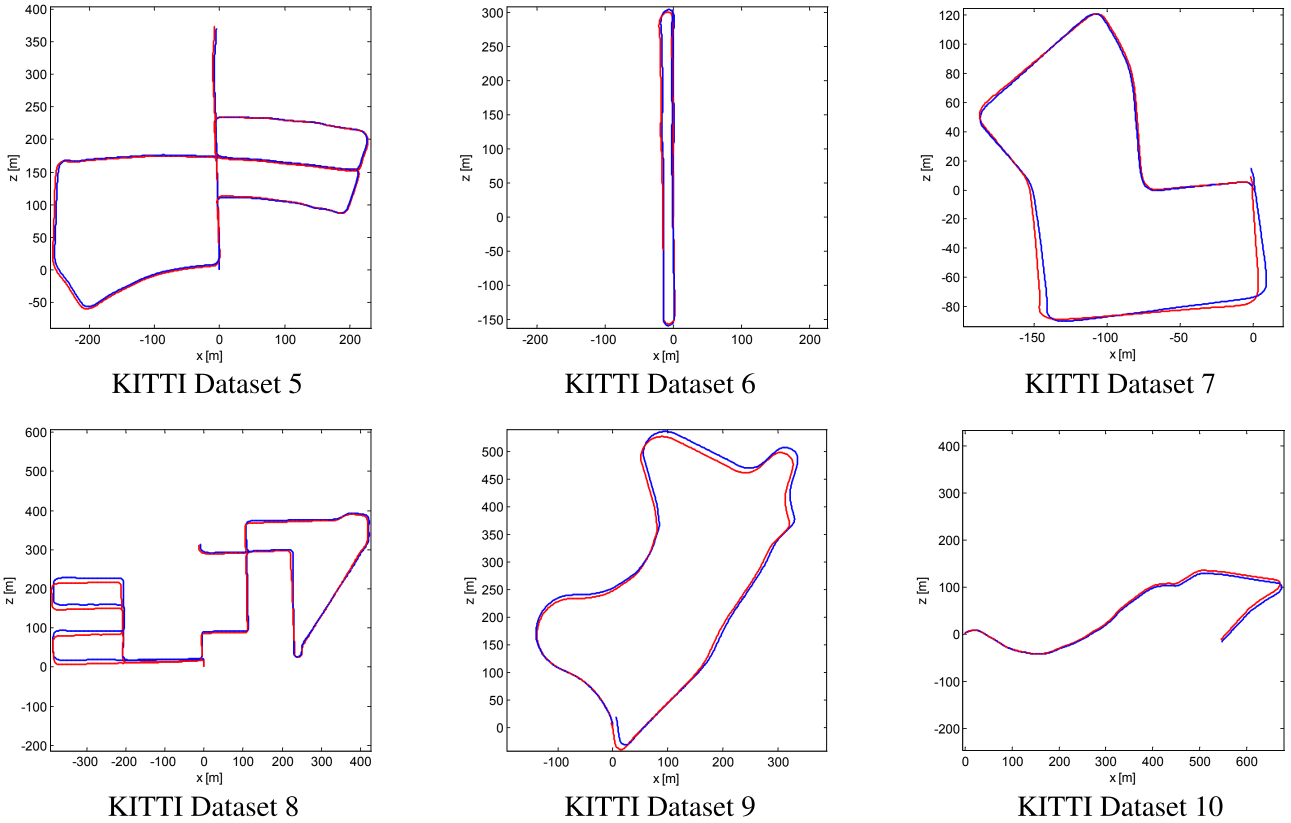
a survival age of a feature is very simple, it partly influences on overall performance in the ego-motion estimation. Moreover, the proposed method is still superior to the baseline algorithm. It demonstrates that the proposed three-point direct method is more accurate than the existing method using all of the inliers that can be partially erroneous.

3.2 Evaluation with test datasets

The results of the proposed method for the test sets of the KITTI dataset are shown in Table 2. The proposed method without additional techniques, Proposed(without MP,U,S), can be compared with VISO2-S [13] and D6DVO [7] which are representative feature-based and direct methods, respectively. Both of them are optimized using many inlier features. Proposed(without MP,U,S) shows the better performance than both the methods. In particular, Proposed(without MP,U,S) is superior to D6DVO which is a direct method using all pixel points in an image. It demonstrates that the motion estimation using a small number of points can provide better performance than the one using lots of features. Here, the performance of Proposed(with MP,U,S) is lower than the performances of other methods [9, 21, 26, 27]. This is because that, while they use the LBA or loop closure that re-optimize camera's poses in key-frames using a bundle of the poses and map data, the proposed method is a frame-by-frame system which is more preferable to the application such as [3]. Its performance can be also more improved applying the LBA or loop closure to the proposed method.

4 Conclusion

In this paper, we proposed the three-point direct stereo visual odometry method. We assume that the ego-motion estimation using a minimum number of accurate features from a contaminated inlier feature set provides a more accurate estimate. Accordingly, we compute ego-motion by sampling only three points and minimizing photometric errors of their local feature patches. The proposed method was evaluated in the publicly available KITTI dataset. The proposed three-point direct method achieved the better performance than the methods that optimize a motion estimate using all inliers. However, the feature positions



KITTI Dataset 5 KITTI Dataset 6 KITTI Dataset 7

KITTI Dataset 8 KITTI Dataset 9 KITTI Dataset 10

Figure 4: Some results of the proposed method. The red and blue lines represent the ground truth and the results of the proposed method, respectively.

reconstructed from stereo images and then updated are not still perfectly accurate. In the future work, we plan to study a feature update method for getting more accurate 3D positions of features and then apply it to the proposed method.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2015R1A2A1A01005455).

References

[1] Hernan Badino and Takeo Kanade. A head-wearable short-baseline stereo system for the simultaneous estimation of structure and motion. In *IAPR Conference on Machine Vision Application*, 2011.

[2] Hernan Badino, Akihiro Yamamoto, and Takeo Kanade. Visual odometry by multi-frame feature integration. In *ICCV*, 2013.

[3] Adrien Bak, Samia Bouchafa, and Didier Aubert. Dynamic objects detection through visual odometry and stereo-vision: a study of inaccuracy and improvement sources. *Machine Vision and Applications*, 25(3):681–697, 2014.

[4] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *IJCV*, 56(3):221–255, 2004.

[5] Fabio Bellavia, Marco Fanfani, Fabio Pazzaglia, and Carlo Colombo. Robust selective stereo slam without loop closure and bundle adjustment. In *ICIAP*, 2013.

[6] Ondrej Chum, Jiri Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition*, pages 236–243. Springer Berlin Heidelberg, 2003.

[7] A.I. Comport, E. Malis, and P. Rives. Accurate quadrifocal tracking for robust 3d visual odometry. In *ICRA*, 2007.

[8] Igor Cvišić and Ivan Petrović. Stereo odometry based on careful feature selection and tracking. In *European Conference on Mobile Robots*, 2015.

[9] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct slam with stereo cameras. In *IROS*, 2015.

[10] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[11] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *ICRA*, pages 15–22, 2014.

[12] Friedrich Fraundorfer and Davide Scaramuzza. Visual odometry: Part ii - matching, robustness, and applications. *IEEE Robotics and Automation Magazine*, 19(2):78–90, 2012.

[13] Andreas Geiger, Julius Ziegler, and Chrstoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium*, 2011.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[15] Robert M. Haralick, Chung-Nan Lee, Karsten Ottenberg, and Michael Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *IJCV*, 13(3):331–356, 1994.

[16] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, 1988.

[17] Jonathan Kelly, Srikanth Saripalli, and GauravS. Sukhatme. Combined visual and inertial navigation for an unmanned aerial vehicle. In Christian Laugier and Roland Siegwart, editors, *Field and Service Robotics*, volume 42 of *Springer Tracts in Advanced Robotics*, pages 255–264. Springer Berlin Heidelberg, 2008.

[18] Ivan Krešo and Siniša Šegvić. Improving the egomotion estimation by correcting the calibration bias. In *10th International Conference on Computer Vision Theory and Applications*, 2015.

[19] S. Li, C. Xu, and M. Xie. A robust o (n) solution to the perspective-n-point problem. *TPAMI*, 34(7):1444–1450, 2012.

[20] Chen-Chung Liu. A global color transfer scheme between images based on multiple regression analysis. *International Journal of Innovative Computing, Information and Control*, 8(1A):167–186, 2012.

[21] Wei Lu, Zhiyu Xiang, and Jilin Liu. High-performance visual odometry with two-stage local binocular ba and gpu. In *IEEE Intelligent Vehicles Symposium*, 2013.

[22] J. M. Martínez Montiel, Javier Civera, and Andrew J. Davison. Unified inverse depth parametrization for monocular slam. In *Robotics: Science and Systems*, 2006.

[23] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *ICPR*, 2006.

[24] David Nister, Oleg Narodisky, and James Bergen. Visual odometry. In *CVPR*, 2004.

[25] Min-Gyu Park and Kuk-Jin Yoon. Efficient point feature tracking based on self-aware distance transform. In *BMVC*, 2012.

[26] Mikael Persson, Tommaso Piccini, Michael Felsberg, and Rudolf Mester. Robust stereo visual odometry from monocular techniques. In *IEEE Intelligent Vehicles Symposium*, 2015.

[27] Taihú Pire, Thomas Fischer, Javier Civera, Pablo De Cristóforis, , and Julio Jacobo Berlles. Stereo parallel tracking and mapping for robot localization. In *IROS*, 2015.

[28] Edward Rosten and Tom Drummond. Machine learning for high speed corner detection. In *ECCV*, 2006.

[29] Martial Sanfourche, Vincent Vittori, and Guy Le Besnerais. evo: a realtime embedded stereo odometry for mav applications. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013.

[30] Davide Scaramuzza and Friedrich Fraundorfer. Visual odometry: Part i - the first 30 years and fundamentals. *IEEE Robotics and Automation Magazine*, 18(4):80–92, 2011.

[31] Carlo Tomasi and Takeo Kanade. *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[32] Zhenhao Zhang and Ying Shan. Incremental motion estimation through modified bundle adjustment. In *ICIP*, volume 2, pages II–343, 2003.