The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

Visual Odometry with Effective Feature Sampling for Untextured Outdoor Environment

Yuya Tamura, Masataka Suzuki, Akira Ishii and Yoji Kuroda
Meiji University, Department of Mechanical Engineering,
1-1-1 Higashimita, Tama-ku, Kawasaki, Kanagawa, Japan
Email:{ce82037, ce92037, ce92003, ykuroda}@isc.meiji.ac.jp

Abstract— **In this paper, we propose stereo vision based visual odometry with an effective feature sampling technique for untextured outdoor environment. In order to extract feature points in untextured condition, we divide an image into some sections and affect suitable processes for each section. This approach can also prevent concentration of feature points, and the influence with a moving object can be reduced. Robust motion estimation is attained using the framework of 3-point algorithm and RANdom SAmple Consensus (RANSAC). Moreover, the accumulation error is reduced by keyframe adjustment. We present and evaluate experimental results for our system in outdoor environment. Proposed visual odometry system can localize the robot's position within 4% error in untextured outdoor environment.**

I. INTRODUCTION

AUTONOMOUS mobile robots have to be able to self-localize while moving in its operational environment. This problem can be solved in various ways using different sensors. In this paper, we use our own stereo camera rig to self-localize. Vision sensor is not only effective to self-localize, but also represent a potential answer to the need of new and improved perception capabilities for mobile robots. The self-position estimation technique using vision sensor is called Visual Odometry (VO). Our aim is that we accomplish self-localization of 6-Degrees-Of-Freedom (6DOF) with VO in any outdoor environments.

In recent years, a number of VO algorithms using either single cameras [1]-[3] or stereo cameras [3]-[7] have been proposed. For instance, in [1], the visual module uses variation of Benedetti and Perona's algorithm [8] for feature detection, and correlation for feature tracking. Robustness is obtained integrating visual data and Inertial Measurement Unit (IMU) by a Kalman filter. However, the monocular algorithm tends to become more complicated than that of stereo visions. Therefore, the monocular algorithm is not appropriate to VO for outdoor environments.

A VO system may lead to fatal errors. The errors are produced by inaccurate feature positions, outliers, the accumulation of errors, and so on. In [3], outlier rejection is achieved using preemptive RANSAC [9]. Because of this, robust visual motion estimation is possible. Konolige et al. in order to reduce the accumulation of errors, bundle adjustment is used [5]. Tomono has applied key frame adjustment instead of bundle adjustment, in order to reduce accumulation of errors and calculation time [7]. These approaches enabled long-distance autonomous movement of a robot. Thus, the effective various robust motion estimation techniques has been proposed and these validity has been proved.

Now, we consider the case where VO is performed in outdoor environments. VO in outdoor environments is more challenging than indoor environments. For example, outdoor environments have various situations where are artificial objects or natural terrains and so on. In artificial environments, the techniques of performing self-position estimation using lines and planes, etc. have been proposed [10][11]. However, utilizing artificial objects like lines and planes is inconvenient for the application in natural terrains. Therefore, these approach are not suitable for VO in outdoor environments. And in [6], odometry provides an estimation of the robot motion that allows limiting a search area for improving feature tracking using a maximum-likelihood formulation for motion computation. This approach is effective in the environments where tires hardly slip. But, in outdoor conditions, robots do not always move in such environments. In addition, since outdoor field might be rough and undulating, localization over such field requires 6DOF motion model.

As mentioned above, there are many problems depending on environment in outdoor circumstances and VO which has adapted itself to all outdoor environment is seldom proposed until now.

In order to perform VO in any outdoor environments, it is very important to solve these problem. So in this paper, the problem of feature points based VO (not using edge or plane) in untextured environments, such as a sand pool, a uniformly white wall, and a paved road is solved. If VO can self-localize of 6DOF (not 3DOF) in untextured circumstances, we believe that it can be utilized in any outdoor environments. That is, we think that VO in the untextured environment is the most difficult problem.

Feature points based VO works by finding interest points and matching them between successive images. That is, the amount of movement is computed using the 3D coordinates of the feature points that is obtained by stereo triangulation before and after movement. However, as previously mentioned, there are some problems in performing VO of the feature point base in untextured environments. An example is shown in Fig. 1. Left image of Fig. 1 shows feature points (red dots) which are extracted using Scale-Invariant Feature Transform (SIFT) algorithm [12] in the image that has been

(a) 50 feature points (b) 400 feature points

Fig. 1. An example of untextured environment. The red circles are the feature points extracted by SIFT.

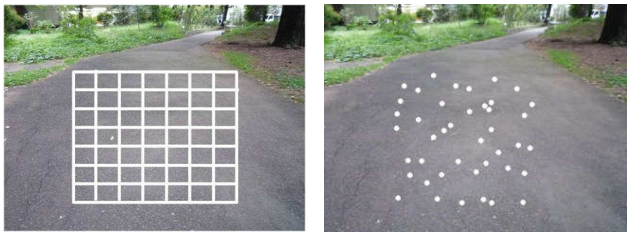


Fig. 2. An example of division of an image. The feature points are extracted from each domain by GoodFeaturesToTrack.

taken on the paved road. As shown in Fig. 1, feature points are hardly extracted from the road-like untextured domain. Many feature points are clustered on distant areas from the camera or the end of the image (e.g. grass). Such untextured region often exists in outdoor environment. There are three problems in performing VO under such circumstance. First, because of the characteristic of a stereo camera, even if these feature points are triangulated, these distance accuracy may be remarkably bad. So most detected feature points in the left image of Fig. 1 are outliers. Next problem is the number of feature points. If the number of the interest points is increased, these can be extracted from the road region. (Right image of Fig. 1) However, this approach is not wise, since the amount of calculation becomes huge and VO system may become unstable. Finally, since Euclidean distances between feature points are not separated enough, estimating motion parameters is difficult.

Then, in this research, we propose a VO with an effective feature sampling technique in consideration of the untextured outdoor environment. The paved road as one example of untextured outdoor environment is considered. In order to extract enough number of feature points scattered on the image in untextured region, we introduce a domain division technique to our VO system. The most robust feature point is compulsorily extracted from each domain. Since, the features tend to cause failure of tracking, robust motion estimation is introduced. The robust motion estimation is achieved by using the framework of three-point algorithm and the framework of random sample consensus. Moreover, the accumulation errors are reduced by key frame adjustment.

In this paper, in order to show the validity of our system, the robot moves around in outdoor environment almost all of which is untextured region. Verification about removal of the outliers of the feature points was performed, the result of the trajectories of proposed VO is compared with ground truth, and the accuracy is verified.

II. ROBUST MOTION ESTIMATION

In this section, we explain the problems of VO in untextured environments and the problems of VO itself.

Then, we explain how these problems are solved. For VO in the untextured environment, a feature points sampling process is an important problem. In order to solve this problem, we propose the feature points sampling process which used area segmentation. Moreover, VO itself has two problems. One of them is outliers. The other is accumulation errors. In order to solve the former problem, we use the framework of RANSAC. In order to solve the latter problem, we use key frame adjustment. Each detail is the following.

A. Feature sampling method

First, we consider the problems of VO in the untextured environments. There are three problems. First, the features cannot be extracted from the untextured domains. Secondly, even if the feature points can be extracted, they tend to concentrate locally. Finally, Euclid distances between such feature points are not separated enough. Such situations make it difficult to presume motion parameters. And, as mentioned above, most detected feature points tend to be selected in distant places from the camera. These feature points might have bad distance accuracy. As a result, fatal errors will be brought to VO.

Then, in this research, in order to solve the above problems, an image is divided in the shape of the grids, as shown in the left image of Fig.2, and the feature point which is the most characteristic in the domain is extracted from each domain (the right image of Fig.2). This approach can also prevent concentration of the feature points simultaneously. Additionally, since the number of the feature points is restricted, increase of calculation cost can be prevented. Thereby, all the problems described previously are solvable.

B. Outlier rejection

Actually, many outliers are contained in the observed feature points. Therefore, many errors arise in the motion estimation. Additionally, if there is a moving object in untextured environment make the problem worse than it already is. It is because almost all the feature point pursues the moving object. Our feature sampling algorithm prevents extracting such the feature point. However, a few outlier is produced. In order to solve this problem, it is necessary to reduce outliers as much as possible. Then, this section describes the technique of removing outliers.

In order to calculate more exact visual odometry, we use the framework of RANSAC and 3-point algorithm. The movement parameter in 6DOF is computable if there are at least 3 set of feature points. Thereby, if N set of feature points are able to track, it will become possible to predict

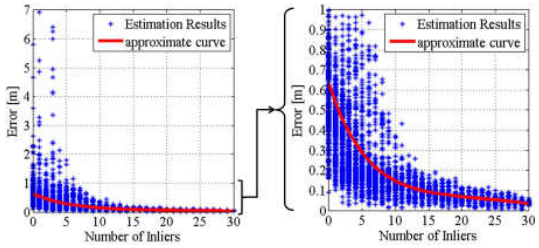


Fig. 3. Relation between error of a motion parameter and number of inliers

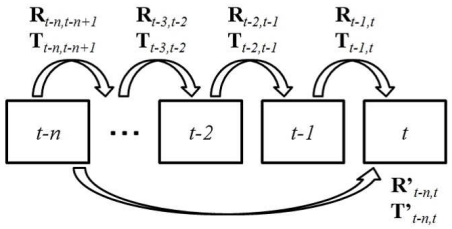


Fig. 4. Key frame adjustment

${}_NC_3$ kinds of motion parameters. In this research, outliers are removed using the framework of RANSAC. The flow which removes outliers is as follows.

1) 3 set of feature points which the Euclid distance left to some extent are extracted out of N set of feature points.
2) A hypothetical movement parameter is computed using these points.
3) All N feature points observed before movement are projected into the current image frame using the hypothesis.
4) We caluculation the reprojection error to each point. If the reprojection error is smaller than a threshold, the data is regarded as inliers.
5) We count the number of inliers.
6) Step1～step5 are repeated M times

Here, the number of times M to repeat is arbitrary in this research. The relation between error of a motion parameter and a number of inliers is shown in Fig. 3. In this figure, the number of max inlier points is 30 (horizontal axis). Vertical axis shows the error of a motion parameter. If there is a number of inliers, it turns out that the errors decrease. Then, outliers are removed using the reprojection error. Moreover, the 3D coordinates of the feature points are updated by averaging them which were computed using the estimated motion parameter and actually measured by triangulation.

C. Key frame adjustment

The motion estimation technique presumes movement sequentially by nonlinear optimization using two frames which continues in time. Thereby, it may lapse into a local minimum in process of optimization. Moreover, the measurment error of feature points causes the error of the motion estimation. The measurement error means that the error by stereo calibration and quantization. Therefore the longer a robot runs a distance , the more accumulation of errors is increased.

In this research, in order to reduce the accumulation error, we use key frame adjustment. That is, a key frame is extracted every several frames. And accumulation of errors are cut down by represuming a camera position between key frames (see Fig.4). Although bundle adjustment is used for the decrease of accumulation errors, calculation cost is high. Then, calculation cost is held down by adjusting only between key frames.

The rotation matrix estimated between time $i-1(i = t-n, \cdots,t)$ and i is set to $R_{i-1,i}$, and a translation vector is set to $T_{i-1,i}$. The motion parameter to $R'_{t-n,t}$ and $T'_{t-n,t}$ can be calculated using equation (1).

$$\begin{cases} R'_{t-n,t} = \prod_{i=t-n}^{t-1} R_{i,i+1} \\ T'_{t-n,t} = \sum_{j=t-n}^{t-1} \left(\left(\prod_{i=t-n}^{j} R_{i,i+1} \right) t_{j,j+1} \right) \end{cases} \quad (1)$$

If the same feature point can be pursuing at least three points in the n past, $R_{t-n,t}$ and $T_{t-n,t}$ are computable. However, since it will be easy to lapse into a local minimum if movement between frames is large, $R'_{t-n,t}$ and $T'_{t-n,t}$ computed by the equation (1) are used as an initial value. A reliable movement parameter is computed using the evaluation function which stated in the foregoing paragraph (equation (3)). $R_{t-n,t}$,$R'_{t-n,t}$, and $T_{t-n,t}$ and $T'_{t-n,t}$ should become the same value theoretically. However this is not completely same. Then, in this research, it asks for $R_{t-n,t}$ and $T_{t-n,t}$ which averaged these two values, and the newest motion parameters $R_{t-n,t}$ and $T_{t-n,t}$ are adjusted.

III. OUR VISUAL ODOMETRY OVERVIEW

Our VO system consists of calibrated stereo camera. Our VO estimates the motion parameter from sequence stereo images. The basic elements of the method are as follows. And the flow chart of our VO is shown in Fig. 5.

A. Camera calibration

Since we are creating the stereo camera uniquely using commercially cheap web cameras, camera calibration and images rectification are two necessary steps in most 3D reconstruction methods using images. To solve the calibration problem, we use Zhang calibration method [13].

B. Feature selection

To extract feature points, we use the Shi-Tomasi feature detector[14]. And, we use our proposed feature points sampling (See Section II).

C. Stereo matching

A stereo matching is used to find correspondences between the extracted feature points between left and right images. The 3D coordinates of each feature point are obtained by triangulation using the feature correspondences. We use Normalized Cross-Correlation (NCC). The window-size is 37×37. In order to reduce the influence of a quantization error, sub-pixel estimation using parabola fitting is used.

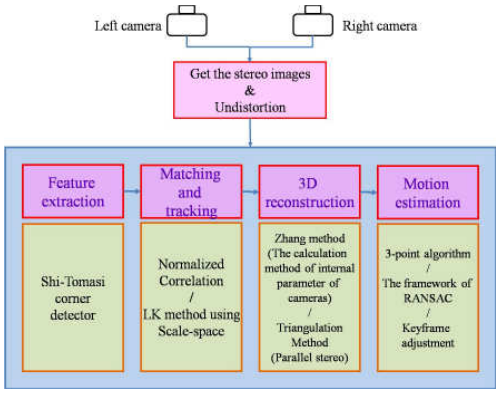


Fig. 5. The flow chart of our VO

D. Feature tracking

The tracking of visual landmarks between consecutive frames is performed by a Lucas-Kanade method using Gaussian pyramid. The Lucas-Kanade method is most suitable for carrying out the tracking of the feature point extracted by the Shi-Tomasi feature detector. By using Gaussian pyramid, it becomes more robust also at large movement of landmarks.In order to reduce the influence of a quantization error, tracking of feature points is performed by the sub-pixel.

E. Motion estimation

The problem of determining the relative motion is solved by two sets of 3D-points $X = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$ and $X' = \{\mathbf{X'}_1, \mathbf{X'}_2, ..., \mathbf{X'}_n\}$. Where $\mathbf{X}_i$ and $\mathbf{X'}_i$ are the i-th 3D-point seen from the robot coordinates before and after motion. In the ideal case, both are changed by the following eqution.

$$\mathbf{X}_i = \mathbf{R}\mathbf{X'}_i + \mathbf{T} \tag{2}$$

Where $\mathbf{R}$ and $\mathbf{T}$ are the rotation matrix and translation vector between each coordinate. However, since there are errors in 3D-points, no points fill a equation (2). The error means that the quantization errors and the error observed by the stereo calibration, etc. Then, we defines the evaluation function shown in equation (3). This is a formula well used in the computer vision community.

$$C = \sum_{i=1}^{N} \|\mathbf{X}_i - (\mathbf{R}\mathbf{X'}_i + \mathbf{T})\|^2 \tag{3}$$

In the equation (3), C is the mean squared error. We compute $\mathbf{R}$ and $\mathbf{t}$ which make C the minimum. Various researches have been studied in order to solve this problem. In this research, a solution based on a *Singular Value Decomposition* (SVD).

IV. EXPERIMENTAL RESULT

In order to verify the validity of the proposed technique, we ran the outdoor autonomous mobile robot which is shown in Fig.6. The experimental field includes the untextured domain. We took datasets from our campus(Data1:587frames, Data2:802frames, Data3:2204frames). The used camera is

Fig. 6. Outdoor mobile root : Infant-Pro & Stereo camera rig

Qcam for Notebooks Pro (QVX-13NS) made from Logicool. The stereo camera is shown in Fig. 6. When observing the about 5 m away place, distance accuracy is $\pm$ 10 cm (the error ratio is 2%). The interface between a camera and a personal computer is USB2.0, and the resolution of an images is VGA (640×480 pix). The stereo image sequences were acquired by about 4 Hz. Moreover, the baseline of the stereo camera is $37cm$ and the depression angle of cameras is 23 degrees. The parameters like the reprojection error used in this experiment are the following.

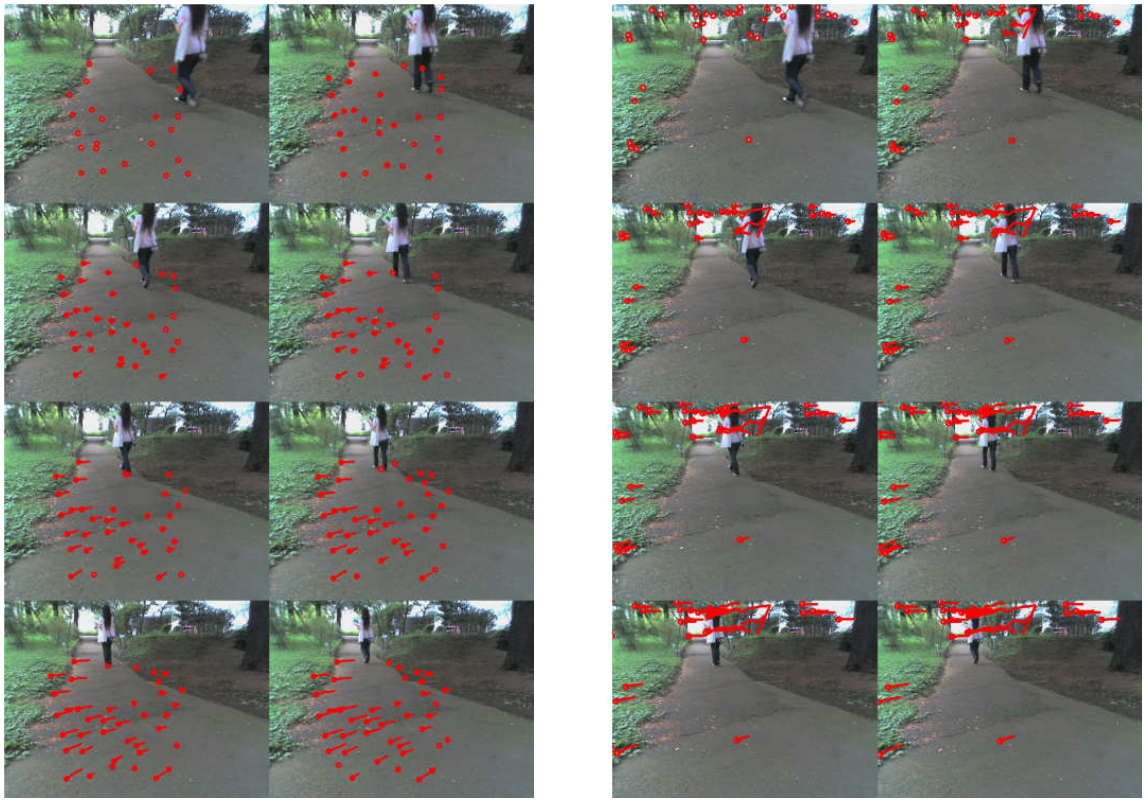
1) The images are divided into the domain of 7×7 as shown in Fig. 2. That is, a maximum of 49 feature points can be acquired.
2) Among these feature points, which reprojection error is less than 1.5 pix are treated as inliers.
3) Three feature points are chosen in case the Euclidean distance between each are at least 0.1m and used to calculate a motion parameter.
4) Convergence calculation will be ended if the number of inliers becomes ten or more points.
5) When the number of inliers is less than ten, the movement parameter obtained before one is used.
6) The five past frames were used for key frame adjustment.

These experiments were done by the system that the CPU is Intel Core2 Duo 2.33 GHz, the RAM is 3.25 GB. OpenCV was used as an image-processing library. The Ground Truth of these experiments is calculated with sensor fusion of Wheel Odometry (WO) and Fiber Optic Gyro (FOG) by using EKF. In the past experiment the robot runs 1km, the error ratio of the system is 1%. This data is enough accuracy to be used as Ground Truth.

A. Verification of our feature points sampling method

The validity of our proposed feature point sampling method was verified. We compared the general technique with our proposed technique using the acquired stereo image sequences. In the general technique, the feature point extraction algorithm used the Shi-Tomasi feature detector and the feature point tracking algorithm used the Lucas-Kanade method using Gaussian pyramid.

The result which performed extraction and tracking of the feature points to the time series data acquired in this experiment is shown in Fig. 7. In Fig. 7, red circles show the extracted feature points, and the red thick lines express the result tracked in several past frames. That is, the feature



(a) Proposed sampling method (b) Standard sampling method

Fig. 7. Extraction and track of the feature points in series image when a pedestrian exists

points with the red thick lines are shown that tracking has succeeded (they are not outliers). It turns out that a moving object (pedestrian) is in Fig. 7. It is important to continue seeing the static feature points for performing exact VO. Therefore, it can be said that the feature points on a moving object are outliers.

In general technique, most feature points are extracted from grasses and trees. Furthermore, the pedestrian causes the miss-tracking of the feature points. Therefore, the calculation of motion parameter diverges in standard approach. On the other hand, our proposed method rejects the outliers when the feature point is newly extracted around a pedestrian, as shown in Fig. 7 (a). In our method, it is assured that there is little influence of outliers and only the reliable feature points are tracked. Moreover, feature points are disperse and would not converge to the moving object since it divides domains and extracts feature points from each domain. Thereby, it is represented that only the static feature points can be extracted even in the environment where a moving object exists. Additionally, using the feature point sampling process which we propose, VO could be calculated stably in the untextured domain where moving object exist.

B. Robot trajectory

In order to prove the availability of proposed method, we performed experiments estimating self position. Although experiment was performed on the paved road, i.e., plane movement, as shown in Fig. 7, VO estimate a 6DOF pose. In order to evaluate various movements, we moved the robot like the blue line in Fig.8. Fig.8 is a result obtained from the experiments. Fig.9 expresses change of a yaw angle for every frame. Both of figures show Data1, Data2, and Data3 in an order from the left picture. Table I shows the the final position errors. Here, although VO is calculated by 6DOF, since the value of Ground Truth and WO are acquired only with 3DOF, Fig.8 is displayed by two dimensions. In the coordinate system of Fig.8, x-axis took the robot advance direction and y-axis just took the robot's leftside. Since experiment was performed on the paved road, tires comparatively seldom slips. Moreover, since it is environment including the untextured domain, this environment is not suitable for VO. Thereby, these experimental environments are disadvantageous situation for VO more than WO overwhelmingly. However, Fig.8 and Fig.9 show that our VO is more accurate than WO in all movements.

In Table I, Fig.8 and Fig.9, the results of 3DOF (x, y, yaw) show that the error of the direction of y is large compared with x direction or a yaw angle. It is considered that the error of the direction of y is caused by the error of the direction of z, and the error of the roll angle from Table.I.

V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed VO with effective feature sampling technique in consideration of the untextured outdoor

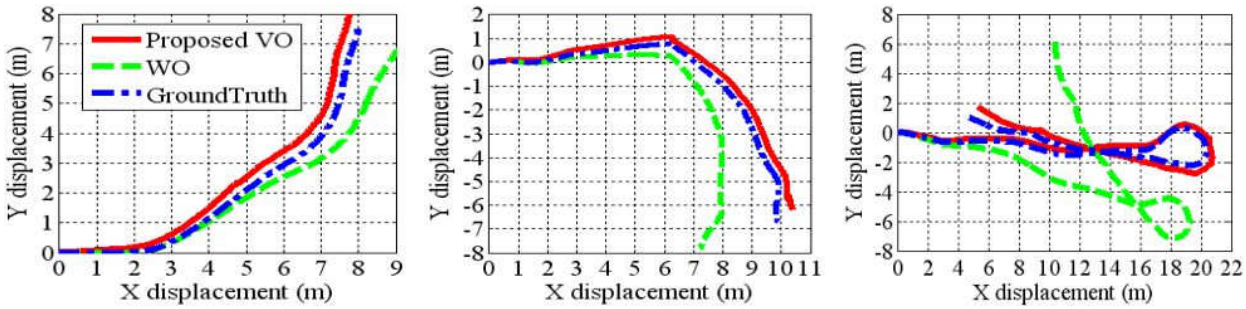


Fig. 8. Estimated robot position

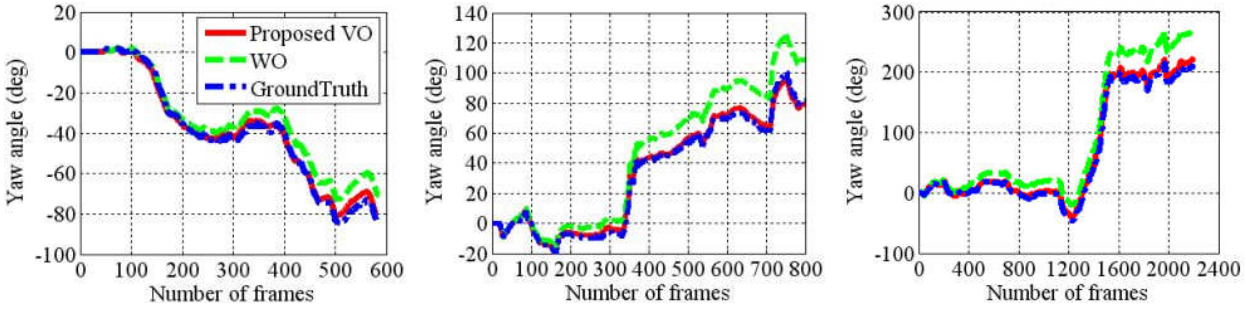


Fig. 9. Estimated robot's yaw angle

TABLE I
COMPARITION OF GROUND TRUTH AND VO

	Data1	Data2	Data3
Error ratio(3DOF)	3.13 (%)	3.05 (%)	2.14 (%)
Error ratio(6DOF)	4.45 (%)	3.08 (%)	2.73 (%)
x	1.12 (%)	4.25 (%)	2.09 (%)
y	6.26 (%)	5.62 (%)	6.57 (%)
z	-0.50 (m)	-0.080 (m)	-0.77 (m)
roll	-0.14 (rad)	-0.23 (rad)	-0.34 (rad)
pitch	-0.003 (rad)	-0.01 (rad)	-0.06 (rad)
yaw	5.60 (%)	3.30 (%)	5.13 (%)

environment. In order to show the validity of our system, the experiment was performed in the outdoor environment including the untextured domain. The experimental results shows that feature points tracking can be stabilized and good estimation of movement parameters can be achieved in the untextured domain. In plane movement, the error over the total amount of movements was within 4%, hence it was proven that highly precise localization was completed. This values of error ratio are the almost same value as the distance accuracy of our own stereo camera rig.

However, the accuracy of roll angle and the direction of the z-axis is not sufficient. Moreover, there is a problem of computationa time (our VO system run at 3-4Hz). Solving these problems is a future subject.

REFERENCES

[1] S. I. Roumeliotis, A. E. Johnson, and J.F. Montgomery, " Augmenting Inertial Navigation with Image-Based Motion Estimation ", *Proceedings of the 2002 IEEE International Conference on Robotics & Automation*, Washington, 2002, pp. 4326-4333.

[2] A.J. Davison, " Real-Time Simultaneous Localization and Mapping with a Single Camera " , *IEEE Int. Conf. on Computer Vision*, Nice, 2003, pp. 1403-1410.

[3] D. Nister, O. Naroditsky, and J. Bergen, " Visual Odometry " ,*Proc. IEEE Computer Society Conference on Computer Vision and Patterrn Recognition*, Volume 1, 2004, pp.652-659.

[4] A. Milella and R. Siegwart, " Stereo-Based Ego-Motion Estimation Using Pixel Tracking and Iterative Closest Point " ,*Proceedings of the Fourth IEEE International Conference on Computer Vision System (ICVS 2006)*, 2006.

[5] K. Konolige, M. Agrawal, and J. Sola, "Large-scale visual odometry for rough terrain", *Proceedings, International Symposium on Research in Robotics (ISRR 07)*, Hiroshima, Japan, 2007.

[6] C.F. Olson, L.H. Matthies, M. Schoppers, and M.W. Maimone, "Rover Navigation Using Stereo Ego-Motion " ,*Robotics and Autonomous Systems*, 43, 2003, pp. 215-229.

[7] M. Tomono, "Robust Stereo SLAM Based on Edge-Point ICP and Error Recovery", *17-th Robotics Symposia*, 2009, pp.217-222

[8] A. Benedetti, P. Perona, " Real-time 2-D Feature Detection on a Reconfigurable Computer " ,*IEEE Conf. Computer Vision and Pattern Recognition*, 1998, pp. 586-593.

[9] D. Nister, " Preemptive RANSAC for Live Structure and Motion Estimation " , *IEEE International Conference on Computer Vision*, Nice, 2003, pp. 199-206

[10] T. Lemaire and S. Lacroix, " Monocular-vision based SLAM using line segments " ,*Proceedings of ICRA 2007*, 2007.

[11] P. Smith, L. Reid, and A. Davison, " Real-Time Monocular SLAM with Straight Lines " , *Proceedings of BMVC2006*, 2006.

[12] S. Se, D. Lowe, and J. Little, "Vision-based Mobile Robot Localization And Mapping using Scale-Invariant Features " , *Proceedings of ICRA 2001*, 2001.

[13] Z. Zhang, "A Flexible New Technique for Camera Calibration", *Pattern Analysis and Machine Intelligence*, vol. 22, 2000, pp. 1330-1334

[14] J. Shi and C. Tomasi, "Good Feature to Track", *IEEE Conference of Computer Vision and Pattern Recognition*, CA, 1994, pp. 593-600