Abstract—Vision sensors are attractive equipments for an autonomous mobile robot because they are information-rich and rarely have restrictions in range and various applications. However, there exists a difficult problem in using vision sensors to reconstruct geometrical information, which is a correspondence or matching problem. We present a simultaneous localization and map building (SLAM) algorithm for an autonomous mobile robot based on a vision sensor robust to the correspondence or matching problem with cooperation of structure from motion (SFM) and stereo. We use an omni-directional stereo vision sensor, which is suitable for SLAM. Due to its large field of view, we can acquire robust estimates and reduce the effect of motion drift from the fiducial coordinate frame. Results on real images illustrate the performance of the proposed method.

I. INTRODUCTION

An autonomous mobile robot must have the ability to navigate at an unknown location in an unknown environment. The simultaneous localization and map building (SLAM) problem have relation to this autonomous ability. SLAM is to build a map by using only relative observation of the environment and to use this map simultaneously to navigate a mobile robot. Although a number of approaches have been proposed to address the SLAM problem using sonar, laser range finder, and millimeter-wave radar (e.g., [1], [2], [3]), vision sensors are attractive equipments for an autonomous mobile robot because they are information-rich and rarely have restrictions in range and various applications. So, the SLAM implemented by a vision sensor would have immeasurable worth in many applications of a mobile robot.

However, there exists a difficult problem in using vision sensors to reconstruct geometrical information, which is a correspondence or matching problem. Many vision based algorithm have suffered from this problem. Among the works recently reported in the filed of SLAM, the system proposed by Davison and Murray [4] is a good advance of the vision based SLAM, which uses an active vision, and shows the vision sensor's attractiveness in real applications. Though this system shows a good performance in unstructured environments, it also has the matching problem of a stereo vision sensor. Especially, the problem will be more difficult when there are inherent ambiguities in matching, such as a repeated pattern over the large part of the scene and occluded region that cannot be seen one of the cameras.

In this paper, we propose a SLAM algorithm based on a vision sensor robust to the correspondence problem with cooperation of structure from motion (SFM) and stereo. SFM is currently one of the most exciting areas of computer vision research. In fact, the purpose of SLAM and SFM is same in that they want to estimate both the motion of the sensor and the shape of the environment as a sensor moves through a scene. Moreover, SFM has the advantage that the correspondence problem is relatively easy to solve. This is due to the fact that many SFM algorithms use multiple images, especially a sequence of images taken at short time intervals [5], [6]. Also, there have been many researches about feature matching or tracking for SFM. Shi and Tomasi [7] proposed a feature selection criterion based on how tracker works and a feature monitoring method that can detect mismatching or mistracking.

Many SFM algorithms based on multiple images can be categorized into two methods. One is a batch method and the other is a on-line method. A batch method determines the estimates using all of the image observations simultaneously and produces most optimal estimates. However, in robot navigation application, each new image must be processed in a short time to produce estimates as it becomes available. This requirement has led to the development of an on-line method. This method has, in general, recursive form and is suitable for use on arbitrary-length image sequences. So, the proposed SLAM exploits an on-line SFM method based on extended Kalman filter (EKF).

In spite of the advantage of easy matching process, SFM has the scale factor ambiguity of monocular vision. An absolute scale cannot be determined without appropriate additional sensors. In many robotic applications, it is not necessary to determine an absolute scale factor, but for autonomous navigation and especially for large-scale SLAM, which is capable of localization and map building over large areas, this scale factor ambiguity must be removed. The proposed method exploits stereo to determine a scale factor. SFM and stereo cooperate to remove the disadvantage of each algorithm. A scale factor ambiguity is removed in a standpoint of SFM and matching problem is solved in a standpoint of stereo even if there are

Abstract—Vision sensors are attractive equipments for an autonomous mobile robot because they are information-rich and rarely have restrictions in range and various applications. However, there exists a difficult problem in using vision sensors to reconstruct geometrical information, which is a correspondence or matching problem. We present a simultaneous localization and map building (SLAM) algorithm for an autonomous mobile robot based on a vision sensor robust to the correspondence or matching problem with cooperation of structure from motion (SFM) and stereo. We use an omni-directional stereo vision sensor, which is suitable for SLAM. Due to its large field of view, we can acquire robust estimates and reduce the effect of motion drift from the fiducial coordinate frame. Results on real images illustrate the performance of the proposed method.

I. INTRODUCTION

An autonomous mobile robot must have the ability to navigate at an unknown location in an unknown environment. The simultaneous localization and map building (SLAM) problem have relation to this autonomous ability. SLAM is to build a map by using only relative observation of the environment and to use this map simultaneously to navigate a mobile robot. Although a number of approaches have been proposed to address the SLAM problem using sonar, laser range finder, and millimeter-wave radar (e.g., [1], [2], [3]), vision sensors are attractive equipments for an autonomous mobile robot because they are information-rich and rarely have restrictions in range and various applications. So, the SLAM implemented by a vision sensor would have immeasurable worth in many applications of a mobile robot.

However, there exists a difficult problem in using vision sensors to reconstruct geometrical information, which is a correspondence or matching problem. Many vision based algorithm have suffered from this problem. Among the works recently reported in the filed of SLAM, the system proposed by Davison and Murray [4] is a good advance of the vision based SLAM, which uses an active vision, and shows the vision sensor's attractiveness in real applications. Though this system shows a good performance in unstructured environments, it also has the matching problem of a stereo vision sensor. Especially, the problem will be more difficult when there are inherent ambiguities in matching, such as a repeated pattern over the large part of the scene and occluded region that cannot be seen one of the cameras.

In this paper, we propose a SLAM algorithm based on a vision sensor robust to the correspondence problem with cooperation of structure from motion (SFM) and stereo. SFM is currently one of the most exciting areas of computer vision research. In fact, the purpose of SLAM and SFM is same in that they want to estimate both the motion of the sensor and the shape of the environment as a sensor moves through a scene. Moreover, SFM has the advantage that the correspondence problem is relatively easy to solve. This is due to the fact that many SFM algorithms use multiple images, especially a sequence of images taken at short time intervals [5], [6]. Also, there have been many researches about feature matching or tracking for SFM. Shi and Tomasi [7] proposed a feature selection criterion based on how tracker works and a feature monitoring method that can detect mismatching or mistracking.

Many SFM algorithms based on multiple images can be categorized into two methods. One is a batch method and the other is a on-line method. A batch method determines the estimates using all of the image observations simultaneously and produces most optimal estimates. However, in robot navigation application, each new image must be processed in a short time to produce estimates as it becomes available. This requirement has led to the development of an on-line method. This method has, in general, recursive form and is suitable for use on arbitrary-length image sequences. So, the proposed SLAM exploits an on-line SFM method based on extended Kalman filter (EKF).

In spite of the advantage of easy matching process, SFM has the scale factor ambiguity of monocular vision. An absolute scale cannot be determined without appropriate additional sensors. In many robotic applications, it is not necessary to determine an absolute scale factor, but for autonomous navigation and especially for large-scale SLAM, which is capable of localization and map building over large areas, this scale factor ambiguity must be removed. The proposed method exploits stereo to determine a scale factor. SFM and stereo cooperate to remove the disadvantage of each algorithm. A scale factor ambiguity is removed in a standpoint of SFM and matching problem is solved in a standpoint of stereo even if there are
ambiguities in matching.

The proposed method adopts two omni-directional vision sensors as a stereo system. Besides producing more robust estimates than conventional cameras, it is suitable vision sensor for SLAM because of its large field of view (FOV). Since the features on image can be retained long time due to the large FOV, the effect of motion drift from the fiducial coordinate frame can be reduced.

II. FEATURES FOR SLAM

It is assumed that a mobile robot moves on the ground-plane that is almost flat and horizontal and the bearings of features are obtained from omni-directional vision sensor. Omni-directional vision sensor consists of a CCD camera and a mirror as shown in Fig. 1(a) and is mounted on the robot so that the axis of symmetry of the mirror is vertical to the ground-plane. The proposed method is targeted at recovering motion and feature locations in a plane by using only the bearings of feature points in the omni-directional image. The proposed SLAM algorithm exploits 2D SFM in static environment. The 1D point features for 2D SFM are extracted from the intensity on horizontal sensing line (HSL) by using 1D edge detection operator. The HSL is determined such that the projections of the points on the HSL constitute horizontal sensing plane (HSP) parallel to the ground-plane. HSL is shown as a closed curve (Fig. 1(b)). The features on HSL are maintained on HSL provided the robot moves on the ground-plane. The two omni-directional vision sensors that form the stereo are aligned such that the HSP of them coincide.

The extracted features are tracked through the image sequences. In our experiments, we used 1D form of Lucas-Kanade tracker [8]. Also, we monitored the quality of image features during tracking by using a measure of feature dissimilarity which is proposed by Shi and Tomasi [7]. Since dissimilarity grows large due to occlusion, disocclusion or other reasons, the feature having large dissimilarity must be discarded from the feature group for SFM.

III. COOPERATIVE METHOD FOR SLAM

In the proposed method which is a two-stage approach, SFM and stereo cooperate to remove the disadvantage of each method. In the first stage, the structure estimates of the scene and motion estimates of the sensor are recovered by using SFM. In the second stage, the scale factor is determined in a stereo matching process by exploiting the recovered results. Consequently, the proposed method reliably recovers the absolute range of the scene and motion parameters while scale ambiguity of SFM and correspondence problem of stereo are removed simultaneously. In the following two subsections, we describe this two-stage approach.

A. SLAM from SFM

The proposed method exploits an on-line SFM method based on EKF. Our SFM method is similar to [5], but replaces the camera model used in that method with the omni-directional camera model of 2D space. The SFM method estimates structure and motion and their covariances. Fig. 2 shows a schematic diagram of the omni-directional vision sensor in the process of observing a feature. A world coordinate frame is defined such that its x and y axes lie in a ground-plane. The three motion parameters of the sensor are represented with a matrix \( R(\theta(t)) \) for rotation and a vector \( t(t) \) for translation, where \( \theta(t) \) is the orientation of the world coordinate frame and \( t(t) \) is the position of the world origin relative to the current sensor coordinate frame.

The vision sensor used in this SFM stage is one of the two omni-directional vision sensors that form the stereo system. Origin of the sensor coordinate frame is on the center of the sensor and y axis is parallel to the horizontal scan line of a camera. A feature reference coordinate frame is assigned to each feature. The feature reference coordinate frame of a feature is the sensor coordinate
frame where that feature is first observed. The rotational parameter \( R_{ri} \) and translational parameter \( t_{ri} \) between the world coordinate frame and the feature reference coordinate frame are saved for each \( i \)th feature and used as constants in the EKF process. As shown in Fig. 2, at the start of the sensor motion, the world and sensor coordinate frame are coincident and the world coordinate frame is the feature reference coordinate frame for the feature observed at the start position.

The position of the \( i \)th feature point, relative to the feature reference frame, is

\[
P_i = (\alpha_i \cos y_{0i}, \alpha_i \sin y_{0i})
\]

where \( \alpha_i \) is the depth and \( y_{0i} \) is the bearing of the \( i \)th feature relative to the feature reference frame.

A feature point relative to the current sensor frame can be represented via

\[
P_{ci}(t) = \begin{bmatrix} X_{ci} \\ Y_{ci} \end{bmatrix} = R(\theta(t))R_{r1}^T P_i - R(\theta(t))R_{r1}^T t_{r1} + t.
\]

Referring to (1) and (2) together, a non-linear measurement equation can be written as

\[
y_i(t) = \arctan(Y_{ci}/X_{ci}) = g(R(\theta(t)), t, \alpha_i, y_{0i}, R_{r1}, t_{r1})
\]

relating the structure of a feature and the motion of a sensor to the current bearing of the feature.

The state vector for the EKF consists of three motion parameters, three velocity parameters, and \( 2N \) structure parameters, where \( N \) is the number of features tracked.

\[
x(t) = (t^T(t), \theta(t), \dot{t}^T(t), \dot{\theta}(t), \alpha_1, y_{01}, \ldots, \alpha_N, y_{0N})
\]

The linear dynamic model is

\[
x(t + \Delta t) = \Phi x(t) + \xi(t)
\]

where

\[
\Phi = \begin{bmatrix} I & 0 & \Delta t I & 0 & 0 \\ 0 & I & 0 & \Delta t I & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{bmatrix}
\]

and

\[
\xi(t) = (\xi^T_1(t), \xi_\theta(t), \xi^T_2(t), \xi_\theta(t), 0^T)^T
\]

is an error term associated with the system modelling and modelled as zero-mean Gaussian white noise sequence. Here, we assume that the robot’s velocity varies very slowly.

The measurement model is a set of \( N \) 1D measurements as follows.

\[
\begin{bmatrix} y_1(t) \\ \vdots \\ y_N(t) \end{bmatrix} = g(x(t), R_{r1}, t_{r1}, \ldots, R_{rN}, t_{rN}) + \eta(t)
\]

(6)

where \( \eta(t) \) is the noise term associated with the bearing measurements and modelled as zero-mean Gaussian white noise sequence. The non-linear vector function \( g(\cdot) \) is vertical accumulation of the scalar function \( g_i(\cdot) \) of (3).

We initialize the state \( \{\alpha_i\} \) all to have an equal value and the states \( \{y_{0i}\} \) to be the bearing measurement of the corresponding features with respect to the feature reference frame. Since the estimates from SFM are subject to arbitrary scaling, we set the scale factor by fixing a static parameter. This fixation is performed by setting the initial variance on an \( \alpha_i \) to zero. Since the EKF implementation by using (5) and (6) is straightforward, the implementation details are not developed further here.

In the proposed method, we have used a polar coordinate, \((\alpha_i, y_{0i})\) rather than a Cartesian coordinate, \((x, y)\) for representing each feature position because the former is more helpful to reduce the dimensionality of the state space. Since the number of the state element is large for large number of features and the observational model is non-linear, the EKF may have unstable and unpredictable estimation behavior due to the initialization. To avoid these results, when starting the EKF algorithm, we set the initial variance on \( y_{0i} \) to zero to fix it at its initial value. This initialization reduces the dimensionality of the state space and lead the EKF to stable results. This trial is reasonable because the uncertainty in \( \alpha_i \) trivializes uncertainty in \( y_{0i} \) and the error variance on \( y_{0i} \) is relatively small. It comes from the fact that the coordinate system used are a polar coordinate system and the features are represented with respect to the feature reference frame rather than the world frame. This fixation on \( y_{0i} \) is similar to that of [5]. However, we do not fix it permanently. When the variance of \( \alpha_i \) (diagonal element of the state covariance matrix)
goes down under a given threshold, we replace the zero variance on $y_{0i}$ with a statistical error variance of bearing observation to estimate more optimal estimates.

In case of the long term navigation, as robot moves, the existing features disappeared gradually and the new features for SLAM must be extracted and tracked. When a feature disappeared, the corresponding state element should be removed from the EKF process. If the $M$th feature disappeared, the removed parameters from the EKF are underlined as follows:

$$x(t) = (t^T(t), \theta(t), \dot{\theta}(t),
\alpha_1, y_{01}, \ldots, \alpha_M, y_{0M}, \ldots, \alpha_N, y_{0N})$$

for $P$ and $Q$

$$K_{2i+1} = \frac{K_{2M+2i} + 5}{K_{2M+2i} + 6},
K_{2i+1} = \frac{K_{2M+2i} + 5}{K_{2M+2i} + 6}$$

for $R$

where $x$ is the state vector, $P$ is the state covariance matrix, $Q$ is $E\{\xi\xi^T\}$ and $R$ is $E\{\eta\eta^T\}$.

When a new feature is extracted and tracked, the new state element corresponding to the feature is inserted into the EKF process as follows:

$$x(t) = (t^T(t), \theta(t), \dot{\theta}(t),
\alpha_1, y_{01}, \ldots, \alpha_N, y_{0N}, \alpha_{N+1}, y_{0N+1})$$

for $P$ and $Q$

In these reforming process for a new feature, we set the initial values of the state vector and state covariance matrix to be the same values that are initially assigned to the features observed at the start position except for the value of state $y_{0i}$. The value of state $y_{0i}$ is set to be the bearing measurement of the corresponding feature at current frame. Since we set the system to search new features after the existing states are converged, the uncertainty estimates of the existing states can be much smaller than that of new features. So, above assignment for the new features do not degenerate the estimates of the other existing states. Moreover, the convergence rate of the new states is faster than that of the existing states observed at the start position due to the interaction with the small uncertainty of the existing states. These effects can be seen in the experimental results of section IV.

B. Scale factor Determination using Stereo Matching

In this section we will elucidate the proposed scale factor determination method using stereo matching algorithm. Since the estimates from SFM algorithm are up to scale, they construct a set of estimates that are spanned with respect to the origin of the world frame. If stereo vision sensor is calibrated, the feature points estimated by SFM relative to one sensor with a scale factor can be re-projected on the other sensor. The scale factor determination method proposed in this paper makes use of an evaluation function that sums the sum of squared difference (SSD) for all these re-projection points. Assume that SFM is performed on the left sensor with a scale factor. The estimated $N$ points $M_i, (i = 1, \ldots, N)$ on the scene relative to current frame are represented by using (7) below.

$$M_i = s (R(\theta)R_i^T \mathbf{P}_i - R(\theta)R_{i0}^T \mathbf{t}_i + \mathbf{t})$$

where $s$ is a scale factor, $R(\theta)$, $\mathbf{t}$ and $\mathbf{P}_i$ are the motion and feature location estimates from SFM.

If the rigid transformation from right sensor frame to left sensor frame are represented with rotation matrix $R_i$ and translation vector $\mathbf{t}_i$, the bearings $m_i, (i = 1, \ldots, N)$, of the re-projected positions on the right sensor are given by

$$M_i = \begin{bmatrix} X_{Mi}^r \\ Y_{Mi}^r \end{bmatrix} = R_i M_i + \mathbf{t}_i$$

and

$$m_i = \arctan \left( \frac{Y_{Mi}^r}{X_{Mi}^r} \right).$$

Finally, we can construct an evaluation function $e(s)$ for matching as follows:

$$e(s) = \sum_{i=1}^{N} \sum_{d \in W} [f(m_i + d) - g(m_i'(s) + d)]^2$$

where, $f(\cdot)$ and $g(\cdot)$ are current image intensity functions on HSL for the left and right sensor respectively, the $\sum_{d \in W}$ means summation over a given feature window $W_i$, and $m_i, (i = 1, \ldots, N)$, are the current bearing measurements of the tracked features on the left sensor.
By searching the scale factor \( s \) that minimizes \( e(s) \) of (10), we can determine the scale factor.

C. Analysis of Cooperative Method

Owing to the proposed matching method using the evaluation function of (10), we can overcome the problem of a false match of a stereo even if there are inherent ambiguities in matching, such as a repeated pattern. In addition, if the number of non-occluded features are larger than that of occluded features (it is true in most case), the proposed method do not suffer from a false match even though some of the interesting features lie on occluded regions in the scene. The advantage of the proposed method is depicted in Fig. 3. It is assumed that the intensity patterns around the feature \( M_1, M_2 \) and \( M_3 \) are same pattern. There is a great possibility of a false match if matching procedure is performed with individual feature. Especially, since the feature \( M_2 \) in occluded region cannot be seen from the right sensor, it is impossible to find a true match. So, false match is to be extracted at all times for feature \( M_2 \). However, if the proposed method is used as shown in Fig. 3, the false match would not occur. For example, even though the feature \( m_2 \) is matched with \( m'_3 \) on a scale factor \( s_3 \), the SSD corresponding to \( m_1 \) and \( m_3 \) in \( e(s) \) would not have minimum value. Consequently, evaluation function \( e(s) \) would not have minimum value on a scale factor \( s_3 \). On the other hand, when \( s_2 \) is selected as a scale factor, \( e(s) \) would have minimum value because \( m_1 \) and \( m_3 \) is matched with \( m'_1 \) and \( m'_3 \) respectively although the SSD corresponding to \( m_2 \) would not have minimum value. Fig. 3 shows that the proposed method can give a robust result against the repeated pattern and occluded regions in the scene.

IV. Experimental Results

To evaluate the location and map building accuracy of the method, we made a controlled environment. The floor was marked with points that should be passed by the system. The distance between the consecutive points is 10mm. The artificial scene features were set up in known positions along the table and wall. To make occluded regions purposely, two pillars were put on the floor. Fig. 4 shows the experimental environment. The omni-directional stereo vision sensor was mounted on a tripod. We took a sequence of 270 images by moving the system via the marked points manually maintaining zero rotation angles. In our omni-directional sensor, the resolution of the 1D image on HSL is 0.2°. The features for SLAM were extracted and tracked by using the Lucas-Kanade tracker and the feature having large dissimilarity was discarded.

Fig. 5 shows the overall experimental results of the proposed SLAM algorithm (see video). The estimated feature positions are designated by cross. The ground-truth feature positions are designated by circles. The dotted lines are plotted from the origin of the feature reference frame to the corresponding feature positions estimated. In this experiment, half of the vertical edges of paper patches were used as features. The estimated sensor path is shown with a solid line and the actual path with a dash-dot line. We can see that there are large errors at the start because the estimates from SFM do not yet converge. If the estimates do not converge, the scale factor also cannot be determined. However, once they converge, this bootstrap period is needed no more. Fig. 6 illustrates the convergence of the structure parameters.

As the sensor moves, some features were mistracked and abandoned by occlusion or disocclusion. The abandoned features are enclosed by a square in Fig. 5 and the
frame number where that features were discarded is shown besides the square. After discarded, the feature positions are updated no more. In this experiment, we set the system to search new features at 150th frame purposely. In fact, the frame should be determined by considering the number of the current states and the distance from the last frame having searched new features. In Fig. 5, the features at bottom-right were newly searched at 150th frame.

Fig. 6 shows that, as described in section III-A, the estimates of the existing states are not degenerated by the reforming process for the new features and the convergence rate of the new states is faster than that of the existing states (for graphical purpose, the feature estimates are set to zero until the features are first observed).

In the procedure of determining a scale factor, we use only the feature states whose variance is under a given threshold because the features not converging yet give rise to a bad effect on the procedure. The first procedure was performed when the number of the converged feature states is larger than a given number. Fig. 7 shows the extracted scale factor as a function of the frame number. In 80th frame, the first scale factor was determined. The scale factor was set to be a random constant until the first procedure was done. In this experiment, we search the scale factor with a resolution of 1mm at every frames and the window size of SSD is 5pixels.

V. CONCLUSION

We have presented a SLAM algorithm for an autonomous mobile robot based on a vision sensor robust to a correspondence problem. We could acquire the method by integrating SFM and stereo. SFM and stereo cooperate to remove the disadvantage of each algorithm. We used two omni-directional vision sensors as a stereo system. As expected, the omni-directional vision sensor gave robust estimates and reduced the effect of motion drift. Of course, feature matching or tracking problem cannot be solved perfectly even in the SFM framework based on multiple images. However, if we abandon the features with high innovation in the measurement step of Kalman filtering, the performance will be maintained. In this paper, the implementation is performed in relatively small and laboratorial environment. It will be validated that the proposed method can be used at large and real environments with an actual mobile robot. Although 1D omni-images of HSL were used in the experiment of this paper, it also must be pointed out that the measurements of the bearings of the features on 2D omni-images can be applied to the proposed method. When 2D images are used, the reconstructed map will be 3D space.

VI. REFERENCES