Abstract - This paper proposes a new algorithm for Simultaneous Localization and Mapping (SLAM) with omnidirectional stereo vision. In our approach, stereo matching is solved efficiently by using the estimated spatial information of the environment and robot motion. The use of two EKF (Extended Kalman Filter) estimators brings about a more reliable robot trajectory and a better map of the environment with more landmarks. The first estimator (a binocular estimator) mostly focuses on the robot trajectory; the second estimator (a monocular estimator) is devoted to the map building. Reliably matched landmarks are entered into the first estimator to establish the reliable robot position and some of the landmark positions. Other landmarks, which have more uncertainty of stereo combination or are observed by only one camera, are passed to the second estimator to establish their positions. This structure results in more precise estimation of robot trajectory and a map of the environment with more landmarks.

Index terms - Omnidirectional vision, SLAM, Stereo matching, Extended Kalman Filter.

I. INTRODUCTION

Vision has played a significant role in SLAM research for nearly two decades. The early idea of SLAM approach was introduced in recursive solutions for Structure From Motion (SFM) [1][2][3][4], where the structure of observed rigid environment and motion of camera are estimated recursively within a sequence of images. These works can be directly applied to SLAM. The vision-based SLAM methods are classified into three main categories: monocular, binocular and mixed methods.

Monocular methods [1][2][3][5][6][7][8] solve the SLAM problem by using a single image sequence captured by one camera attached to a mobile robot. The advantage of this approach is the ease of tracking and matching in a single sequence of video. Assuming a slow robot speed, trackers such as Kanade-Lucas-Tomasi feature tracker [9] keep the correspondence in a single video sequence effectively. However, monocular methods usually lack a true scaling factor that results in the loss of real position and cannot take advantage of the direct triangulation.

Binocular methods [10][11] use two different viewpoint images captured by stereo camera. The advantage of a binocular method is the direct triangulation at a single stereo observation. The baseline of the stereo camera can be set up large enough for the environment to result in the reliable recovery of the observed landmarks. Hence, binocular methods can overcome the problem of scaling factor and can give the direct position of observed points. However, they face the stereo correspondence between video sequences. In the researches [10][11], the authors used isolated stereo matching with a current stereo pair of images. These approaches are hard in a dynamic environment and a waste of calculation, because some matched features are re-matched in the following sensor readings, and there is always the same risk of mismatching.

Mixed monocular and binocular methods have existed for the better stereo matching solutions in SLAM. Molton et al. proposed stereo SFM [4]. In this method, they used multiple frames to estimate the motion of the camera and solve the stereo matching. However, this approach is not robust enough because only four frames -two latest frames from the left sequence and two latest frames from the right sequence- are considered at the same time. There are also papers in computer vision for multi-frame matching [12][13][14], which give much more reliable results of the matching and are more robust in dynamic scenes. These are most optimal for stereo matching problems. However, they are batch algorithms that are not suitable for our real-time SLAM problem. In another approach to a mixed method, Kim and Chung [15] used stereo matching to solve the scaling factor problem of the previous SFM [1][2]. Their stereo matching solution was helped by the estimated landmarks from SLAM, and this brought about a better result than the conventional matching of features without the spatial relations. However, this upgrade of binocular-based SFM did not have the advantage of stereo vision in the estimation. Another disadvantage of these stereo methods was the restriction in the number of estimated landmarks and the reliabilities of all the landmarks varied.

There are some factors that influence the estimation. First, the differences of stereo observing angles vary widely. Second, the noise on observations of landmarks is different. Third, when the environment is dynamic, there are moving features points. Finally, there are also errors that come from the tracking and extraction of features. Therefore, the reliabilities of the features are inconsistent. If using all feature points at once, the overall performance of estimation can be contaminated. In current papers on vision-based SLAM, this problem is either not mentioned [4][11] or not fully mentioned [10]. Garcia and Solanas [10] proposed a SLAM method based on the estimation of robot ego-motion from the motion of feature points on the image plane after solving the stereo matching by current stereo observation. The outlier ego-motions from the dominant ego-motion were deleted with the related feature points. However, a fewer number of landmarks is estimated in single stereo pair of frames.

In this paper, we propose a stereo vision-based method using an omnidirectional sensor and Extended Kalman filter.
The omnidirectional image sensor has a large field of view (FOV) and rich information can be extracted via the sensor. A number of researchers [5][7][11][15] also took that advantage and developed SLAM methods for robotics applications. In our method, stereo matching is done with help from SLAM. The matching uses the spatial information estimated from SLAM for searching stereo pairs on two sequences of image frames. It is done by recursively calculating the uncertainty of all possible combinations and selecting one with a small value as a true combination. There are two types of estimators in our method: the binocular estimator to deal with robot position and landmarks that have the reliable stereo correspondence and the monocular estimator to estimate the positions of remaining landmarks that cannot give reliable estimates from the first estimator.

The contributions of our paper are twofold. Firstly, in our method stereo matching is done efficiently and simultaneously with SLAM and secondly, the structure of the two estimators (binocular estimator, monocular estimator) results in more reliable robot trajectory, landmark positions and more landmarks being estimated. Moreover, this algorithm can also work under dynamic environment with moving objects.

II. OVERVIEW OF THE METHOD

Like most previous vision-based methods, we assume the slow velocity of the robot, both translational velocity and rotational velocity. This helps the feature tracker work effectively in each video sequence and reduces the linearization error in EKF. In the current work, our algorithm is designed to work with a 2D map of the dynamic environment and 2D motion (translation and rotation on planar ground) of the robot. In practice it can be applied to an office environment where there are plenty of vertical lines. These static features are considered as the landmarks of an environment and are included in the map.

We define a stereo combination as the combination of one feature on the first camera and one feature on the second camera. To use the stereo information, we must know the correspondence of features. In our algorithm, we track a combination by time using an uncertainty value and decide its reliable correspondence when the value becomes less than a threshold. This method of correspondence tracking falls into a multi-hypotheses tracking so that it can improve the stereo correspondence. When new features appear, there is large number of stereo combinations related to them, however by the time robot moves the uncertainty values of these combinations become distinguished, the bad combinations are eliminated. The most reliable stereo combinations among the remaining are then selected into the binocular estimator to estimate the robot motion. The other features which are not reliably given a stereo correspondence are passed into the monocular estimator. At the initialization, binocular estimator can not estimate the robot motion because we can not use the uncertainty value yet. However, in this period, we can use dead-reckoning with the odometry sensor to keep the robot motion to update the uncertainty of the combinations. After few frames, binocular estimator can run with the reliable stereo combinations and our system begins. The overview of our algorithm is described in Fig. 1.

In this paper, we assume that the robot has two omnidirectional vision sensors, camera 1 and camera 2, and moves (translates and rotates) on the 2D ground. Translational velocity \( v \) and rotational velocity \( \omega \) are given by odometry sensors on the robot. The robot is located at \((x_R,y_R,\phi)\), which coincides with the position of camera 1, in a global coordinate system, as shown in Fig. 2. \( d \) is the baseline between the two cameras.

In the omnidirectional stereo vision, 2D information of a landmark \( L \) can be described by the pair of observing angles (right observing angle \( \alpha_{i2} \) and left observing angle \( \alpha_{i1} \)). Using triangulation, we can recover the position of the landmark as follows:

\[
x_{L_{i}}[k] = x_{R}[k] + \frac{d * \sin(\alpha_{i2}[k])}{\sin(\alpha_{i2}[k] - \alpha_{i1}[k])} \cos(\alpha_{i1}[k] + \phi[k]),
\]

\[
y_{L_{i}}[k] = y_{R}[k] + \frac{d * \sin(\alpha_{i2}[k])}{\sin(\alpha_{i2}[k] - \alpha_{i1}[k])} \sin(\alpha_{i1}[k] + \phi[k]),
\]

where all measurements are taken at time \( k \).
As the robot moves, if we know the robot position, the position of each stereo combination is estimated by (1) and (2). For each combination \( L_i \), we call the set of \( n \) continuous position estimations \( S(L_i,n) \). Then statistical calculation on \( S(L_i,n) \) will give us the uncertainty of this combination and then show if it is reliable or not.

IV. STEREO CORRESPONDENCE AND RELIABLE LANDMARKS

We use an evaluation value, the observation angle uncertainty, for finding a true stereo combination. An estimated landmark position of a true combination should be stable through a recursive estimation process. Whereas the wrong or combination of moving features causes unstable estimations. We describe the instability as angular uncertainty of stereo combination and it is shown in Fig. 3.

For each stereo combination, at each observation by time, we can estimate its position. After a short sequence of observations, we get an historical collection of \( n \) positions \( S(L_i,n) \). Angular direction of each position \((x_{L_i}[k-j], y_{L_i}[k-j])\) from the current robot position, at time \( k \) is defined as:

\[
\beta_{i}[k-j] = \arctan\frac{y_{L_i}[k-j] - \hat{y}_{R}[k]}{x_{L_i}[k-j] - \hat{x}_{R}[k]}, \quad (j = 0,1,\ldots,n-1),
\]

where, \( \hat{X}_R[k] = \hat{X}_R[k-1] + \delta V[k-1] \) robot position \( X_R[k] = [x_R[k], y_R[k], \phi[k]] \) and velocity vector \( V[k] \) from odometry.

The standard deviation of the direction \( \beta_i \) is derived as:

\[
\text{angdev}(L_i) = \left(\frac{1}{n-1} \sum_{j=0}^{n-1} (\beta_i[k-j] - \beta_i)^2\right)^{1/2}.
\]

This standard deviation is used as the angular uncertainty.

There are some observations from statistics for this angular uncertainty of stereo combination:

+ **Observation 1**: Estimated positions of true combination from a static landmark are slightly varied around the true position due to noise and error in robot position. The uncertainty for this combination is small.

+ **Observation 2**: Estimated positions of wrong combination vary largely due to the movement of the robot and also error in robot position. The uncertainty for this combination is relatively large and unbounded.

+ **Observation 3**: Estimated positions of true combination of moving features are varied due to the movement of the feature and the error in robot position. The uncertainty for this combination is also large but bounded.

![Fig. 3. Calculation of angular uncertainty](image)

![Fig. 4. Position distribution of wrong and true combinations](image)

Fig. 4 shows the image of the observation of the wrong and true combinations. The distribution region for the true combination is smaller than the distribution region for the wrong combination. Therefore, we can identify the reliably true correspondence of static feature points by comparing the uncertainty. A threshold is set to get the reliable correspondence.

V. TWO ESTIMATOR STRUCTURE

Our algorithm takes advantage of two estimators because of the target-oriented design. The binocular estimator is assigned to focus on the localization of the robot. The monocular estimator is devoted to building the map of static landmarks, where the map is as large as possible. They must cooperate to achieve the objective of Simultaneous Localization and Map building.

In the next sub-sections, we only summarize the implementations of binocular and monocular estimator. The details can be derived from [16][17].

A. Binocular estimator

Dynamic state vector \( X[k] = [x_R[k], y_R[k], \phi[k]] \) of this estimator consists of robot position \( X_R[k] \) and the reliable landmarks \( L_i, X_{L_i}[k] = [x_{L_i}[k], y_{L_i}[k]] \).

1) Prediction

In our algorithm, we use odometry to predict the robot position. Translational velocity \( v \) and the rotational velocity \( \omega \) are measured by the encoders. \( v \) and \( \omega \) are assumed to be contaminated by zero-mean white Gaussian noise processes \( \nu \) and \( \nu_\omega \). The covariance matrices for the noise processes are given by \( Q[k] = E[\nu[k] \nu[k]^T] \), and \( N_\omega = E[\nu_\omega[k] \nu_\omega[k]^T] \). The propagation of the state vector is described as follows:

\[
\hat{x}_R[k] = \hat{x}_R[k-1] + \delta t \cos(\phi[k-1]) v[k-1],
\]

\[
\hat{y}_R[k] = \hat{y}_R[k-1] - \delta t \sin(\phi[k-1]) v[k-1],
\]

where \( \hat{x}_R[k] \) and \( \hat{y}_R[k] \) denote the estimated robot position. The covariance matrices for the state vector is given by:

\[
R[k] = R[k-1] + Q[k].
\]

With the reliable landmarks, the sensory data from each camera can be estimated as:

\[
\hat{x}_{L_i}[k] = \hat{x}_{L_i}[k-1] + \delta t \cos(\phi[k-1]) v[k-1],
\]

\[
\hat{y}_{L_i}[k] = \hat{y}_{L_i}[k-1] - \delta t \sin(\phi[k-1]) v[k-1],
\]

where \( \hat{x}_{L_i}[k] \) and \( \hat{y}_{L_i}[k] \) denote the estimated landmark positions. The covariance matrices for the sensory data is given by:

\[
S[k] = S[k-1] + R[k].
\]
B. Monocular estimator

The propagation of the state vector is not linear and we can use a Taylor expansion to linearize the transition.

2) Correction

Every time robot observes the environment via a stereo camera the feature points are tracked by a tracker and identified. Therefore, we know the correspondence in each single sequence of images. These sensor readings are used to correct the prediction in the above phase. Observation for each selected combination \( L_i \) is the stereo pair of observations \((\alpha_{i1}[k], \alpha_{i2}[k])\). Therefore, observation vector is

\[
Y[k] = [\alpha_{i1}[k] \alpha_{i2}[k] \ldots \alpha_{Ni1}[k] \alpha_{Ni2}[k]]^T.
\]

In our algorithm, the triangulation functions (1) and (2) are used as the observation functions for each stereo combination. We rewrite (1) and (2) as follows

\[
\dot{x}_L[i][k] = \ddot{x}_R[i][k] + f_{ii},
\]

\[
\dot{\dot{x}}_L[i][k] = \ddot{x}_R[i][k] + f_{ii},
\]

(9)

(10)

where,

\[
f_{ii}(\alpha_{i1}[k], \alpha_{i2}[k], \dot{\phi}[k]) = \frac{d}{\sin \alpha_{i1}[k] - \alpha_{i1}[k]} \left( \frac{\cos \alpha_{i2}[k] + \dot{\phi}[k]}{\sin \alpha_{i1}[k] - \alpha_{i1}[k]} \right), \quad (11)
\]

\[
f_{ii}(\alpha_{i1}[k], \alpha_{i2}[k], \dot{\phi}[k]) = \frac{d}{\sin \alpha_{i1}[k] - \alpha_{i1}[k]} \left( \frac{\sin \alpha_{i2}[k] + \dot{\phi}[k]}{\sin \alpha_{i1}[k] - \alpha_{i1}[k]} \right). \quad (12)
\]

Combining the observation functions (9) and (10) for all \( N \) selected combinations, we get a system of nonlinear equations. To use the Extended Kalman Filter, the equations (11) and (12) must be linearized.

3) Select reliable landmarks for binocular estimator

As stated in Section II and IV, the angular uncertainty from current robot position for combination is used to select the reliable stereo combinations as the static reliable landmarks. By doing this we also solve the stereo matching for some of all features. However, the stereo matching is done recursively in a multi-frame manner. When new stereo pairs come in, we update the uncertainty and select new reliable combinations of features.

B. Monocular estimator

The features which are not selected in the binocular estimator come from unreliable stereo pairs, landmarks observed by only one of the two cameras, and moving object features. The monocular estimator is designed to deal with all these features except the moving object features, which can not be the landmarks. Although both cameras can be used in this estimator, only one of the two cameras is selected for the task, such as camera one.

Although we can use the SFM algorithm for this monocular estimator, we take advantage of the above stereo estimator because the monocular estimator can utilize the robot position which is known from the binocular estimator. The virtual stereo camera is set up as follows: the “first camera” is camera one at the time of the first appearance \( t \). The “second camera” is at current time \( k \); and the baseline is the Euclidean distance \( d(X_0[t], X_k[k]) \). A single Extended Kalman Filter is set up for the state vector that consists of positions of all these static landmarks.

1) Prediction

The prediction of the state vector for landmarks is itself, where

\[
\dot{X}_L[i][k] = \ddot{X}_L[i][k - 1].
\]

This is linear function, which does not need to be linearized

2) Correction

Applying the triangulation for this landmark point \( L_i \), we get

\[
\dot{x}_L[i][k] = \ddot{x}_R[i][k] + g_{i1},
\]

\[
\dot{\dot{x}}_L[i][k] = \ddot{x}_R[i][k] + g_{i1},
\]

(14)

(15)

where

\[
g_{i1}(\alpha_{i1}[t], \alpha_{i2}[k], \dot{\phi}[t]) = d\dot{X}_R[i][k] \times \frac{\sin(\alpha_{i1}[t] - \dot{\phi}[t]) + \dot{\phi}[t]}{\sin(\alpha_{i1}[t] - \dot{\phi}[t]) + \dot{\phi}[t]} \right), \quad (16)
\]

\[
g_{i2}(\alpha_{i1}[t], \alpha_{i2}[k], \dot{\phi}[t]) = d\dot{X}_R[i][k] \times \frac{\sin(\alpha_{i2}[k] - \dot{\phi}[t]) + \dot{\phi}[t]}{\sin(\alpha_{i2}[k] - \dot{\phi}[t]) + \dot{\phi}[t]} \right), \quad (17)
\]

and \( \Phi[k] = \arctan(\frac{\dot{x}_R[i][k] - \dot{\phi}[t]}{\dot{\phi}[t] - \dot{\phi}[t]}). \)

The observation functions of the monocular estimator are linearized similarly to those of binocular estimator. There is one note for this estimator as following: because we use \( X_R[i][t], X_R[k] \) as constant in this estimator we should consider the additional uncertainties in the covariance matrix of transition noise. In other words, the landmark position “moves” a little with the error of the robot trajectory.

3) Static landmarks

As we know from Section IV, the uncertainty of the moving features is greater than the uncertainty of static landmarks. We can apply the same technique used to select landmarks for binocular estimator to select the static landmarks for monocular estimator.

VI. EXPERIMENTS

A. Simulation experiments

Experiments were carried out by various simulations and were compared to the SLAM solution presented in [15], which is an upgrade to omnidirectional vision of previous EKF-SFM methods [1][2]. The first experiment was to compare the performance of the proposed method and the EKF-SFM method. The second was carried out in the simulated dynamic environment. And then stereo matching was examined in those above results.

The simulated environment consists of 59 static landmarks, and they only appeared at the distance less than 14[m] to the current camera location. The robot moved at the translational speed \( v = 0.05[m/s] \), and the rotational velocity slowly changed in the sinusoid function \( o[k] = 0.05-0.0015 \)
sin(0.0314k)[Rad/s]. Omnidirectional stereo camera has the baseline \(d = 0.5\text{[m]}\). The robot odometry was distorted by the white Gaussian noise; the standard deviations for translational velocity noise and rotational velocity noise were \((0.032\text{[m/s]}, 0.0065\text{[Rad/s]})\), respectively. EKF-SFM was supplied with the true scaling factor at the initialization. The simulations were carried out with 800 iterations of sensor readings.

1) Static environment

Fig. 5. Estimation of proposed method.

Fig. 5 depicts an example of estimation experiments of our proposed method with simulated environment. The simulated observation error with \(Dev = 0.0026\text{[Rad]}\) was added for both cameras. In this figure, the small circles denote the true landmarks; the black stars are the estimated landmarks, they are combined outputs from two estimators. The red trajectory denotes the ground-truth trajectory of the robot; the dark blue trajectory is the estimated trajectory. The comparison with EKF-SFM is shown in Fig. 6 and Fig. 7.

![Fig. 6. RMS errors of estimated robot position, robot orientation and landmark positions by proposed method (red line) and EKF-SFM (dark blue dash line).](image)

Fig. 6 shows the error of estimated robot position, robot orientation and landmark positions by proposed method (red line) and EKF-SFM (dark blue dash line).

2) Dynamic environment

Experiment with dynamic environments was carried out by adding some moving features with the speed similar to robot speed to the above static environment. The result showed that combinations related to these features were ignored from binocular estimator by their large angular uncertainty, and these features were also discarded from the monocular estimator for the same reason. The experiment was not different from above static environment.

3. Stereo matching

In the experiments, our method solves the stereo matching efficiently. Fig. 8 shows the result of the experiment as shown in Fig. 5. There is no wrong combination selected. At the beginning there are a lot of possible combinations. However over time with the update of uncertainties, the completely wrong ones were discarded from the next consideration. The number of correctly selected landmarks is very close to the number of real landmarks.

B. Real experiments

Real experiments were carried out with a Pioneer 2-DXe robot of ActivMedia Robotics. This mobile robot was attached with an omnidirectional stereo camera with baseline 226[mm] and connected to a PC (company: HP, product name: xw6000, CPU: 2.8GHz, memory: 1GB) through RS232C communication ports. The environment was the corridor of our building and landmarks were the vertical edges. The actual robot motion and landmark positions were measured and used as the ground-truth. In real experiments, occlusion and mistracking problems usually occurs. In our case, estimation was carried out with 350 video frames and there existed mistracking for about every 25 frames, such as exchange of...
identifying numbers between extracted features or same feature with different identifying numbers when its track was lost. In our algorithm, mistracking makes the uncertainty of related combinations become large and these combinations would be ignored after more few frames. If the observation noise covariance matrix is set reasonably large enough, the EKF still can work. Fig. 9 and Fig. 10, show the estimation our proposed method and also EKF-SFM method, in which the legends are similar to those of Fig. 5. Table I and Table II show the detailed comparison of the average results for both methods after three running times. We again meet the same results as shown in simulation results.

In current program, although we have not focused on the improving the computational cost, the total processing time for each frame is about 0.12[sec] with appearance of about 30 detected features. It is reasonable to be applied in real environment.

VII. CONCLUSIONS AND FUTURE WORKS

In this paper we have proposed a new robust method for SLAM using a stereo omnidirectional vision sensor. In our method stereo matching is performed by using spatial information estimated by the binocular estimator. Two combined estimators are designed for different purposes. The first estimator focuses on the reliable localization; meanwhile the second is devoted to the map building. This idea is not restricted to binocular vision. For example, if a robot has a trinocular sensor, it is easy to expand to three combined estimators; trinocular, binocular, monocular estimators. We used an uncertainty measurement of the combinations for deciding the true stereo pair. This value is derived simultaneously from the deviation of estimated positions from the binocular estimator.

Experiments showed our stereo matching method was very effective. Convergence speed and accuracy were shown to be much better than the EKF-SFM method. It can also work under a dynamic environment where some moving objects appeared. Our proposed algorithm can discriminate the static landmarks, the false matches and also moving object points. However, we simply delete the moving object points. Further work can be added to track these points.

REFERENCES