The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Stereo-based Simultaneous Localization, Mapping and Moving Object Tracking

Kuen-Han Lin and Chieh-Chih Wang

Abstract— **Vision based simultaneous localization and mapping (SLAM) has recently received much research interest. However, vision based SLAM could be corrupted with the inclusion of moving entities, which makes it hard to operate in dynamic environments. Simultaneous localization, mapping and moving object tracking (SLAMMOT) serves as a solution to deal with moving objects while performing SLAM. The existing work has shown the feasibility of monocular SLAMMOT in dynamic environments. However, monocular SLAMMOT inherits the observability issue of bearings-only tracking in which moving entities would be unobservable according to motions of the camera and moving objects. In this paper, stereo-based SLAMMOT is proposed to solve the observability issue as well as increase the accuracy of localization, mapping and tracking. Simulation and experimental results demonstrate that the proposed stereo SLAMMOT is superior than monocular SLAMMOT in dynamic environments.**

I. Introduction

Over the past twenty years, simultaneous localization and mapping (SLAM) has acquired extensive research interest in the robotics literature. It serves as a basic component for robots exploring in unknown environments. Promising SLAM has been demonstrated with the use of laser scanners to achieve large scale and accurate mapping [1]. Recently, the trend has been moving to using cameras to perform SLAM for their light weight and low-cost features. In addition, rich appearance and texture information of the surroundings is also available to the users. The first remarkable work in visual SLAM was done by Davison *et al.* [2] in which a single camera is used to perform SLAM under the extended Kalman filtering (EKF) framework. Montiel *et al.* [3] proposed an inverse depth parametrization for monocular SLAM. This parametrization initializes features with no delay and successfully estimates both near and distant features. Due to bearing only information is available, monocular SLAM only reconstructs the environment up to a scale if no other information (e.g. odometry of camera) or prior (e.g. features depth) is available. Therefore, stereo vision systems are often applied to avoid this scale ambiguity in which feature depths can be directly estimated. Stereo-based SLAM has been demonstrated in indoor environments [4] [5]. Sola *et al.* [6] further pointed out that fusing monocular information using the inverse depth parametrization from stereo cameras increases the estimability for far features and proposed the BiCamSLAM. Paz *et al.* [7] demonstrated 6 degrees of freedom stereo-based SLAM by combing the inverse depth parametrization and Euclidean parametrization. Although vision-based SLAM has successfully demonstrated in large scale outdoor environment, most SLAM works assume the environment is static which could be impractical for robots operating in dynamic environments [8] [9].

Wang *et al.* [10] proposed a theoretical framework to solve the simultaneous localization, mapping and moving object tracking (SLAMMOT) problem and demonstrated the feasibility of SLAMMOT using laser scanners. There are a few attempts to accomplish vision-based SLAM and tracking in dynamic environment. Sola [11] proposed BiCamSLAM with a rule-based moving object detection method. Tracking was done separately and individually for each moving object in a robocentric representation. Migliore *et al.* [12] proposed monocular SLAM and moving object tracking in which moving objects are detected by a simple statistic test and tracked by separated bearing only trackers. However, both works separate SLAM and tracking for reducing the computational complexity. Both works mentioned the observability issue in bearing only tracking(BOT) also occurs in monocular SLAMMOT in which the camera needs to perform higher order trajectories to estimate the full state of moving objects [13]. Wang *et al.* [14] recently proposed an augmented state approach to monocular SLAMMOT in which the states of moving objects are augmented into the SLAM state vector, and demonstrated that moving object tracking could improve the SLAM performance. To avoid the unobservable conditions, the camera performed spiral motions for achieving converged tracking results.

This work is an extension work of augmented state SLAMMOT using a single camera. We proposed a stereo-based SLAMMOT approach to overcome the observability issue. Two cameras are treated as two observers, and measurement updates are performed for each instance in EKF. Monte Carlo simulations and real experimental results show that stereo SLAMMOT is outperformed monocular SLAMMOT in terms of the performance of SLAM and tracking and the observability issue.

The rest of paper is organized as follows: The theoretical foundation of monocular SLAMMOT are briefly reviewed in Section II. In Section III, the proposed stereo SLAMMOT is introduced and the simulation results are described. Section IV shows the experimental results using real image sequences and Section V addresses our conclusion and future work.

K.-H. Lin and C.-C. Wang are with the Department of Computer Science and Information Engineering and the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan e-mail: linsm@robotics.csie.ntu.edu.tw, bobwang@ntu.edu.tw

978-1-4244-6676-4/10/$25.00 ©2010 IEEE

II. Theoretical Foundation

In this section, we described the theoretical foundation of monocular SLAMMOT proposed by Wang *et al.* [14] for understanding the proposed stereo SLAMMOT approach described in the next section.

A. Monocular SLAMMOT using the Inverse Depth Parametrization

In monocular SLAMMOT, the camera location $\mathbf{x}_k$, map features $\mathbf{m}_k^i$ and moving objects $\mathbf{o}_k^i$ are simultaneously estimated using the EKF algorithm. The states of moving objects are augmented into the original SLAM state vector to improve the SLAM and moving object tracking performance. SLAMMOT state vector χ are defined as:

$$\chi = (\mathbf{x}_k^\top, \mathbf{m}_k^{1\top}, \mathbf{m}_k^{2\top}, \ldots, \mathbf{m}_k^{q\top}, \mathbf{o}_k^{1\top}, \mathbf{o}_k^{2\top}, \ldots, \mathbf{o}_k^{n\top})^\top \tag{1}$$

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^c & \mathbf{q}_k^c & \mathbf{v}_k^c & \omega_k^c \end{pmatrix}^\top \tag{2}$$

where $\mathbf{x}_k^c$ and quaternion $\mathbf{q}_k^c$ with the norm constraint $\| \mathbf{q}_k^c \| = 1$ [15] denote the camera location and orientation in the world coordinate system. $\mathbf{v}_k^c$, ω_k^c represent the velocity and the angular velocity of the camera, respectively. Both map features and moving objects are coded using the inverse depth parametrization. The velocities of moving objects in the global coordinate are also coded. In this work, the constant velocity model is applied to describe the motions of moving objects. The states of $\mathbf{m}_k^i$ and $\mathbf{o}_k^i$ are:

$$\mathbf{m}_k^i = \begin{pmatrix} x & y & z & \theta & \phi & \rho \end{pmatrix}^\top \tag{3}$$

$$\begin{aligned} \mathbf{o}_k^i &= \begin{pmatrix} x_k & y_k & z_k & \theta_k & \phi_k & \rho_k & v_k^x & v_k^y & v_k^z \end{pmatrix}^\top \\ &= \begin{pmatrix} \mathbf{p}_k{}^\top & \mathbf{v}_k{}^\top \end{pmatrix}^\top \end{aligned} \tag{4}$$

where $(x \quad y \quad z)^\top$ represents the camera optical center position with respect to the world coordinate system when the feature was first observed. θ and ϕ are azimuth and elevation of the ray defined in the world coordinate. ρ is the inverse depth between the feature and camera optical center. $\mathbf{p}_k$ represents the moving object position in the inverse depth parametrization and $\mathbf{v}_k$ is the velocity estimate of moving object in the world coordinate. Let $\mathbf{r} = (x \quad y \quad z)^\top$ and the direction vector $\mathcal{G}(\theta, \phi)$ defines the direction of the ray, then the point can be transformed from inverse depth parametrization to 3D points as:

$$\mathbf{r} + \frac{1}{\rho}\mathcal{G}(\theta, \phi) \tag{5}$$

In the prediction statge of EKF, camera motion model is assumed to be constant velocity (CV) and constant angular velocity (CAV). For motion model of moving objects, CV model is applied. Thus the equations in the prediction stage of EKF are:

$$\mathbf{x}_{k+1} = \begin{pmatrix} \mathbf{x}_k^c + \mathbf{v}_k^c \Delta t \\ \mathbf{q}_k^c \otimes q(\omega_k^c \Delta t) \\ \mathbf{v}_k^c \\ \omega_k^c \end{pmatrix} \tag{6}$$

$$\begin{aligned} \mathbf{p}_{k+1} &= \mathbf{p}_k + \mathbf{v}_k \cdot \Delta t \\ &= \mathbf{r} + \frac{1}{\rho_k}\mathcal{G}_k + \mathbf{v}_k \cdot \Delta t \\ &= \mathbf{r} + \frac{1}{\rho_{k+1}}\mathcal{G}_{k+1} \end{aligned} \tag{7}$$

where $\mathcal{G}_{k+1}$ and ρ_{k+1} are the new direction vector and inverse depth of moving objects.

In the measurement stage of EKF, Both map features and moving objects are transformed to the camera coordinate system and then projected on the camera image plane. Let R_k^c be the rotation matrix defined by the camera orientation $\mathbf{q}_k^c$. The points are transformed to the camera coordinate by:

$$\begin{aligned} \mathbf{h}_k^{m_i} &= \begin{pmatrix} h_x^{m_i} \\ h_y^{m_i} \\ h_z^{m_i} \end{pmatrix} = h(\mathbf{m}_k^i, \mathbf{x}_k) \\ &= \mathrm{R}_k^c \left(\mathbf{r} + \frac{1}{\rho}\mathcal{G}(\theta, \phi) - \mathbf{x}_k^c \right) \end{aligned} \tag{8}$$

$$\begin{aligned} \mathbf{h}_k^{o_i} &= \begin{pmatrix} h_x^{o_i} \\ h_y^{o_i} \\ h_z^{o_i} \end{pmatrix} = h(\mathbf{o}_k^i, \mathbf{x}_k) = h(\mathbf{p}_k^i, \mathbf{x}_k) \\ &= \mathrm{R}_k^c \left(\mathbf{r} + \frac{1}{\rho_k}\mathcal{G}_k(\theta, \phi) - \mathbf{x}_k^c \right) \end{aligned} \tag{9}$$

The predicted measurements on the image plane are:

$$\mathbf{z}_k^{m_i} = \begin{pmatrix} u \\ v \end{pmatrix} = Proj(\mathbf{h}_k^{m_i}) = \begin{pmatrix} u_0 - \frac{f}{d_x}\frac{h_x^{m_i}}{h_z^{m_i}} \\ v_0 - \frac{f}{d_y}\frac{h_y^{m_i}}{h_z^{m_i}} \end{pmatrix} \tag{10}$$

$$\mathbf{z}_k^{o_i} = \begin{pmatrix} u \\ v \end{pmatrix} = Proj(\mathbf{h}_k^{o_i}) = \begin{pmatrix} u_0 - \frac{f}{d_x}\frac{h_x^{o_i}}{h_z^{o_i}} \\ v_0 - \frac{f}{d_y}\frac{h_y^{o_i}}{h_z^{o_i}} \end{pmatrix} \tag{11}$$

where $Proj$ is the project function, (u_0, v_0) is the camera center in pixels, f is the focal length, d_x and d_y represent the pixel size. The monocular SLAMMOT state vector is updated by the EKF algorithm.

B. Moving Object Detection in Monocular SLAMMOT

In the monocular SLAMMOT approach, it is assumed that objects can be classified as stationary or moving. The detection method is based on an observation that misclassifying a moving object and adding this moving object into SLAM would significantly degrade the SLAM performance. Both the camera estimate and stationary object estimates would be affected. Therefore, moving objects could be detected by examining monocular SLAM results under different hypotheses. Two local monocular SLAMs are initialized under two hypotheses when a new feature is extracted. One is local monocular SLAM without adding this new feature and the other is local monocular SLAM under the assumption

that this new feature is stationary. The difference of two hypotheses is defined as:

$$d_m = (\mathbf{x}_m^c - \mathbf{x}_r^c)^T \mathbf{\Sigma}_{\mathbf{x}_r^c}^{-1} (\mathbf{x}_m^c - \mathbf{x}_r^c) \tag{12}$$

where $\mathbf{x}_r^c$ and $\mathbf{x}_m^c$ are camera poses for two hypotheses without and with adding this new feature. $\mathbf{\Sigma}_{\mathbf{x}_r^c}$ is the covariance matrix of $\mathbf{x}_r^c$. Instead of making a hard decision to classify the feature type, the differences of two hypotheses are temporally integrated using a binary Bayes filter. By setting up an inverse measurement model $p(\mathcal{H}_m|d_m)$ properly, the log odds ratio can be computed as:

$$l_t(\mathcal{H}_m) = log\frac{p(\mathcal{H}_m|d_m)}{1-p(\mathcal{H}_m|d_m)} - log\frac{p(\mathcal{H}_m)}{1-p(\mathcal{H}_m)} + l_{t-1}(\mathcal{H}_m)$$

Since d_m is a a chi-square variable, $p(\mathcal{H}_m|d_m)$ is similar to chi-square distribution. After a fixed number of updates, a new feature could be classified as static if the log odds ratio $l_t(\mathcal{H}_m)$ is larger then a predetermined threshold λ, otherwise the feature is classified as moving.

In addition, adding misclassified moving objects into the state vector of monocular SLAM often induce unexpected negative inverse depth estimates. Negative inverse depths could originate from inconsistencies between measurements and predictions, which is likely to happen if a moving object is assumed as stationary. Therefore, the appearance of negative inverse depths is also checked for detecting moving objects. Both binary Bayes filter method and negative inverse depths check are integrated by a decision tree. The detection method first examines the inverse depth of the new feature in local SLAM, if inverse depth is still positive, the binary Bayes filter method is then performed.

III. Stereo SLAMMOT

In this section, the detail of the proposed stereo SLAMMOT approach is described. The effect of stereo SLAMMOT to the observability issue is discussed. The comparisons of monocular SLAMMOT and stereo SLAMMOT are shown by Monte Carlo simulations.

A. Stereo SLAMMOT

The proposed stereo SLAMMOT follows the monocular SLAMMOT framework with additional measurement update from the second camera. The two cameras are calibrated and images are rectified with the both principle axes pointing to the Z direction in the camera coordinate. The stereo camera setup and the camera coordinate system used in this work are shown in Fig.1. The right camera represents the origin of the camera coordinate and the left camera is placed along the x axis with the baseline vector $\mathbf{b} = (b \quad 0 \quad 0)^\top$ where b is the distance between two cameras. To project the point into the left camera, the point's relative position to the left camera is:

$$\mathbf{h}_{k,l}^y = \mathbf{h}_{k,r}^y - \mathbf{b} \tag{13}$$

where $\mathbf{y} \in \{\mathbf{m}_k^{1\top}, \mathbf{m}_k^{2\top}, \ldots, \mathbf{m}_k^{q\top}, \mathbf{o}_k^{1\top}, \mathbf{o}_k^{2\top}, \ldots, \mathbf{o}_k^{n\top}\}$ is the point state vector of map features or moving objects. The projections in the left camera are $\mathbf{z}_{k,l}^y = Proj(\mathbf{h}_{k,l}^y)$.

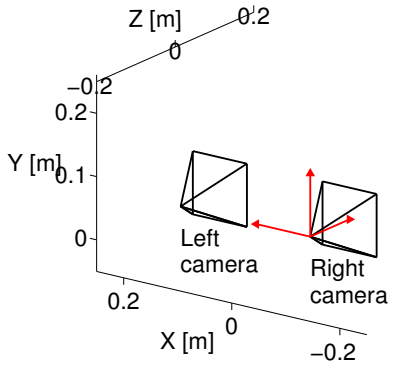


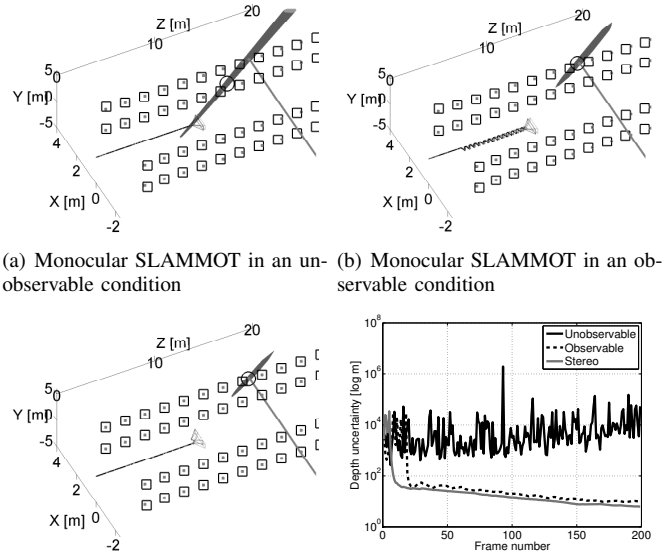
Fig. 1. The coordinate system of the stereo cameras. The origin is set on the optical center of the right camera. The left camera is set along the x axis with the distance of baseline.

Since $\mathbf{h}_{k,r}$ and $\mathbf{h}_{k,l}$ is only different by a baseline vector, Jacobians of two measurements are similar in which the only difference is the projection Jacobians. Given the Jacobians of the measurements of the two cameras, the SLAMMOT state vector is updated accordingly.

In order to deal with features at far distances, the proposed method makes the use of the inverse depth parametrization. As shown by Paz *et al.* [7], the inverse depth parametrization has more correct uncertainty modelling for features at far distances. The XYZ parametrization tends to overestimate features around near regions and underestimate features at far distances.

B. Observability of Stereo SLAMMOT

It has been shown that multiple observers can avoid the unobservable conditions in bearings-only tracking. Fig.2(a) and Fig.2(b) show the monocular SLAMMOT results in an unobservable condition and in an observable condition, respectively. In the unobservable condition, the camera moves forward with a speed of $0.5m/s$. In the observable condition, a circular motion on the XY surface is added in the constant velocity motion when the moving object appears in the view of the camera. It is clearly shown that the state of the tracked moving object can not be converged in the unobservable condition. Fig.2(c) shows the stereo SLAMMOT result in the same condition of Fig.2(a). The stereo camera follows the constant velocity motion with the baseline of 24cm. In both monocular SLAMMOT and stereo SLAMMOT, the depth's 95% uncertainty bound is set to cover from a predefined close distance of 2.1 meters to infinity in the inverse parametrization. Fig. 2(d) shows the moving object depth uncertainty under three conditions. The largest principle axis of the 95% uncertainty volume is computed and the range of uncertainty volume along the principle axis is defined as the depth uncertainty in this work. Under the unobservable condition, the depth uncertainty is large ($10^3 \sim 10^4$ meters). In the observable and stereo conditions, the depth uncertainty quickly decreased and converged below 10 meters after 150 frames. The result showed that stereo SLAMMOT does not have the observability issue.



(a) Monocular SLAMMOT in an unobservable condition

(b) Monocular SLAMMOT in an observable condition

(c) Stereo SLAMMOT in an unobservable condition

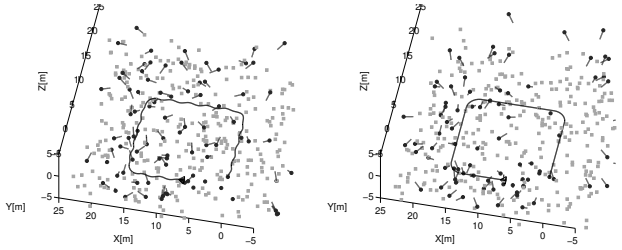
(d) Depth Uncertainty Comparison

Fig. 2. Observability in monocular SLAMMOT and stereo SLAMMOT. Grey dots and grey lines represent the true locations of static objects and the true trajectories of the moving object, respectively. Light grey line indicate the camera trajectory ground truth. Squares and circles indicate the location estimates of the static objects and the moving object, respectively. Only the moving object uncertainty is drawn for clarity.

C. Comparisons of Monocular SLAMMOT and Stereo SLAMMOT

The comparisons of monocular SLAMMOT and stereo SLAMMOT are evaluated with 40 Monte Carlo simulations in total. The simulated camera has a resolution of 320×240 and a focal length of 170 in pixel unit. This setup makes the simulated camera have a wide field-of-view of 86 degrees. In stereo SLAMMOT, the stereo camera have a baseline of 24 cm. Fig. 3 shows the simulation scenarios for both cases. Map features are randomly located within a rectangular cube with a width of 30 m and a height of 10 m. Moving objects are randomly placed in the camera's field-of-view with a maximum depth of 10 m. Each moving object moves with a constant speed of 0.75 m/s and a random initial moving direction. There are about 50 moving objects and 140 map features in each Monte Carlo simulation. The synthetic images are generated by the projections of map features and moving objects into camera. Zero mean Gaussian image noise with σ of 1 pixels are added in each image. In the monocular SLAMMOT simulations, the camera moves spirally to avoid the observability issue with a trajectory length of 59 m. In the stereo SLAMMOT simulations, camera moves at a constant velocity with a trajectory of 56 m.

Table I summaries the detection results of the two approaches. Both approaches have a similar detection rate of 0.8 on detecting moving objects and the false alarm rate of 0.1 on misclassifying map features as moving. The detection performance is not affected in the stereo SLAMMOT approach. For evaluating SLAM and moving object tracking



(a) Monocular SLAMMOT

(b) Stereo SLAMMOT

Fig. 3. The Monte Carlo simulation environments. The black solid line indicates the camera trajectory. Grey squares and black circles indicate true stationary and moving objects, respectively. The trajectories of moving objects are also shown.

performance, the root mean square error (RMSE) of the estimated camera trajectories and moving objects positions in 40 Monte Carlo simulations are used. Moving objects positions are evaluated in camera coordinate for decoupling SLAM performance and tracking performance. Table II shows that stereo SLAMMOT improved the accuracy of localization and moving object tracking. Unlike monocular SLAMMOT needs known feature depths at the beginning to estimate the scale, stereo SLAMMOT can be automatically operated in dynamic environments and provides more accurate SLAMMOT estimates than monocular SLAMMOT.

TABLE I
DETECTION RESULT

	True moving	False static	True static	False moving
Stereo	896	224	2552	290
Monocular	927	186	2597	271

TABLE II
SLAMMOT PERFORMANCE

	Camera RMSE	Moving object RMSE
Stereo	**0.17 m**	**0.30 m**
Monocular	0.71 m	1.12 m

IV. EXPERIMENTAL RESULTS

In this section, the proposed stereo SLAMMOT approach is evaluated using real image sequences. The limited ground truth is provided by laser based SLAMMOT on the XZ plane.

A. Experiment Setting

Fig. 4 shows the robotic platform, NTU-PAL7, in which a Point Grey Bumblebee X3 stereo camera was used to collect image data and a SICK S200 laser scanner was used for ground truthing. The stereo cameras have a baseline of 24 cm and a field-of-view of 66 degrees. Image data are acquired at 10 Hz and the resolution of the collected images is 640×480. Fig.5 shows the experiment environment of a corridor scenario (2.7 m $\times$ 14.5 m). Three pedestrians with their trajectories are shown in Fig.5. The first pedestrian went

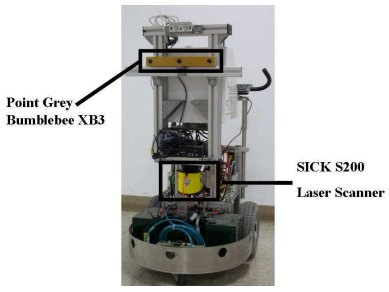
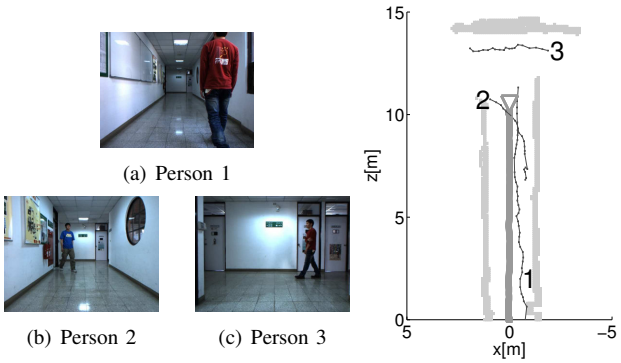


Fig. 4. The NTU PAL7 robot.



(a) Person 1

(b) Person 2

(c) Person 3

(d) LIDAR-based SLAMMOT

Fig. 5. The ground truth of the real data experiment from LIDAR-based SLAMMOT. The trajectories of the camera and three pedestrians are denoted as thick and thin grey lines, respectively. The map ground truth is drawn with light grey color.

toward the positive z direction and disappeared around the corner (frame 53 to 200). The second pedestrian went across the corridor and toward the camera (frame 286 to 342). The third pedestrian went along the positive x direction (frame 440 to 507).

Image features extraction were done using the good feature approach [16], and then a non-maximum suppression is applied to obtain a sparse feature set from images. The selected features were tracked by the Kanade-Lucas-Tomasi (KLT) tracker [17]. In cases that the KLT tracker fails, an active feature [18] searching method is used. After obtaining the image feature tracking result on first camera, the images features on second camera were obtained directly using established correspondences from the stereo camera.

B. Experimental Results

In this experiment, the camera moved at a constant velocity. Table. III shows the moving object detection results. As there were more than one selected features on each person, 21 correct moving features were detected by our detection method. The tracking performance was evaluated on those successfully detected moving objects. 6 map features which misclassified as moving were discarded in evaluation.

Fig. 6 shows the stereo SLAMMOT result at three instants. Map features corresponding to the ceiling or the floor were removed manually for emphasizing moving object tracking results. It is showed that moving objects were successfully detected and tracked by the proposed approach. Even in the condition that Person 1 appeared in the image sequence for long time, our SLAM result was not corrupted in dynamic environments. The depth uncertainty of these three persons went below 1 meters when the true depths were within 3.5 meters. In the case that moving objects left away from the camera, the depth uncertainty became larger as the distance estimates were more uncertain at far distances.

Table IV shows the camera localization RMSE and the tracking RMSE over the three person trajectories. The result showed that our approach achieved a low localization error of 0.40 m and low tracking error with average of 0.46 m. Person 2 had a higher position error as the 2D image tracking was temporally lost. Relocating the target caused a higher location error. The performances of localization, mapping and tracking all verify that the proposed stereo SLAMMOT approach is feasible in dynamic environments.

TABLE III
DETECTION RESULT

	Ground Truth	
	Moving objects	Map features
Moving	26	6
Stationary	4	128

TABLE IV
LOCALIZATION AND TRACKING POSITION ERROR

	Camera	Trajectory		
		1	2	3
RMSE	0.40 m	0.29 m	0.49 m	0.46 m

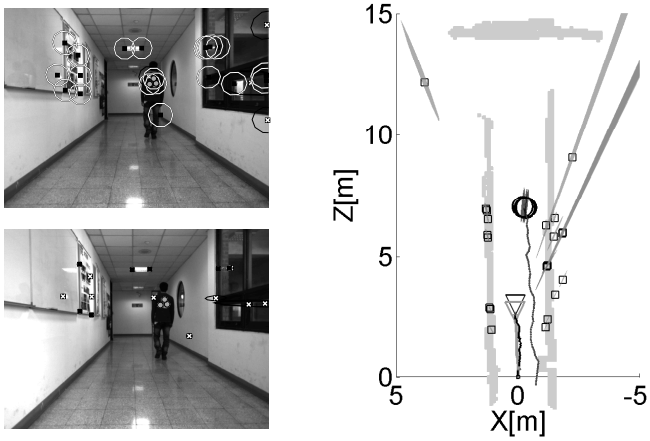
V. CONCLUSIONS AND FUTURE WORKS

In this paper we proposed the stereo SLAMMOT approach to avoid unobservable conditions and demonstrated the promising tracking results in both simulations and the real experiment. With the use of the inverse depth parametrization, we performed EKF updates in original monocular SLAMMOT for two cameras. No special camera trajectory is needed to perform tracking. The Monte Carlo simulations showed the overall performance of stereo SLAMMOT is superior than monocular SLAMMOT and the real experiment results demonstrated the feasibility of the proposed stereo SLAMMOT approach in dynamic environments.

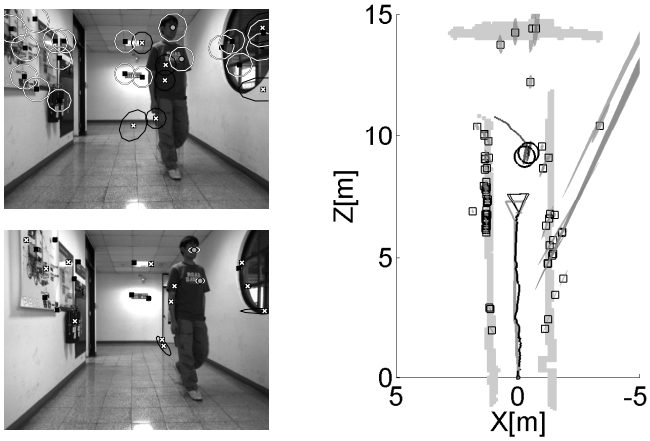
One limitation of this work is the need of reliable 2D feature tracking over images to perform stable stereo SLAMMOT. It has been discussed to model static features as planar patches to assist feature tracking in vision based SLAM [19]. However, it could be more challenging in the moving object cases because of sudden appearance changes. In the future we plan to investigate the use of more dense image features [20] to increase the stability of 2D feature tracking on moving objects. In addition, the stereo SLAMMOT approach could serve as a guideline to decide which region should have denser features according to the SLAMMOT estimates.

REFERENCES

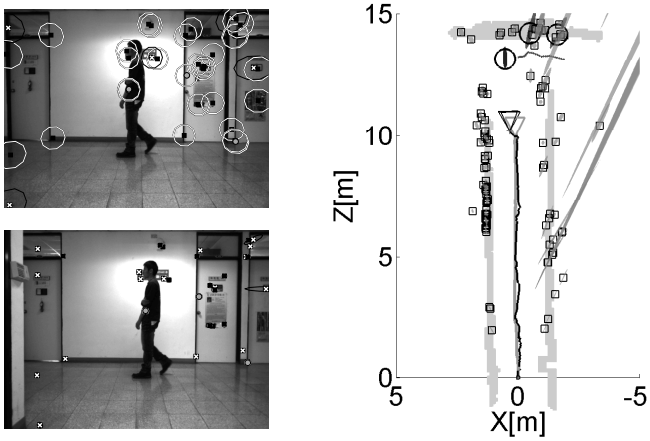
[1] D. M. Cole and P. M. Newman, "Using laser range data for 3d slam in outdoor environments," in *Proceedings of the IEEE International*



(a) Person 1



(b) Person 2



(c) Person 3

Fig. 6. The experimental results. (a)(b)(c) showed the stereo image pairs and the results at frame 136, 315 and 468. The upper image is the right camera view and the lower image is the left camera view. In the image, Black solid squares and grey circles indicate map features and moving objects with associated new measurement and being updated. The white ellipses show the projected 2σ bounds of the estimates of the map features and moving objects. White cross with its black uncertainty bound indicate map features without any associated measurement and not being update. The result showed three moving persons were tracked with a small depth uncertainty.

Conference on Robotics and Automation, (ICRA) , Orlando, Florida, USA, 2006, pp. 1556–1563.

[2] A. Davison, "Real-Time Simultaneous Localisation and Mapping with a Single Camera," in *IEEE International Conference on Computer Vision*, September 2003, pp. 1403–1410.

[3] J. M. M. Montiel, J. Civera, and A. J. Davison, "Unified inverse depth parametrization for monocular slam," in *Robotics: Science and Systems*, Philadelphia, USA, August 2006.

[4] S. Se, D. Lowe, and J. Little, "Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks," *International Journal of Robotics Research*, vol. 21, pp. 735–758, 2002.

[5] R. Sim, P. Elinas, and M. Griffin, "Vision-based slam using the rao-blackwellised particle filter," in *Proc. International Joint Conference on Artificial Intelligence (IJCAI) Workshop on Reasoning with Uncertainty in Robotics(RUR)*, 2005.

[6] J. Solà, A. Monin, and M. Devy, "Bicamslam: Two times mono is more than stereo," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Roma, Italy, April 2007, pp. 4795–4800.

[7] L. M. Paz, P. Pinies, J. D. Tardos, and J. Neira, "Large scale 6 dof slam with stereo-in-hand," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 946–957, October 2008.

[8] C.-C. Wang and C. Thorpe, "Simultaneous localization and mapping with detection and tracking of moving objects," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Washington, DC, May 2002.

[9] S. Wangsiripitak and D. W. Murray, "Avoiding moving outliers in visual slam by tracking moving objects," in *IEEE International Conference on Robotics and Automation (ICRA)*, Kobe, Japan, May 2009, pp. 375–380.

[10] C.-C. Wang, C. Thorpe, S. Thrun, M. Hebert, and H. Durrant-Whyte, "Simultaneous localization, mapping and moving object tracking," *The International Journal of Robotics Research*, vol. 26, no. 9, pp. 889–916, September 2007.

[11] J. Sola, "Towards visual localization, mapping and moving objects tracking by a mobile robot: a geometric and probabilistic approach." Ph.D. dissertation, Institut National Polytechnique de Toulouse, February 2007. [Online]. Available: http://homepages.laas.fr/jsola/JoanSola/eng/JoanSola.html

[12] D. Migliore, R. Rigamonti, D. Marzorati, M. Matteucci, and D. G. Sorrenti, "Use a single camera for simultaneous localization and mapping with mobile object tracking in dynamic environments," in *ICRA Workshop on Safe navigation in open and dynamic environments: Application to autonomous vehicles*, 2009.

[13] E. Fogel and M. Gavish, "Nth-order dynamics target observability from angle measurements," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 3, pp. 305–308, May 1988.

[14] C.-C. Wang, K.-C. Wang, C.-H. Hsiao, K.-H. Lin, and Y.-L. Chao, "Monocular vision-based simultaneous localization, mapping and moving object tracking," *submitted to a journal*, under review.

[15] R. Zanetti, M. Majji, R. H. Bishop, and D. Mortari, "Norm-constrained kalman filtering," *Journal of Guidance, Control, and Dynamics*, vol. 32, no. 5, pp. 1458 – 1465, 2009.

[16] J. Shi and C. Tomasi, "Good features to track." in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[17] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 1981, pp. 674–679.

[18] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, June 2007.

[19] N. D. Molton, A. J. Davison, and I. D. Reid, "Locally planar patch features for real-time structure from motion," in *Proc. British Machine Vision Conference(BMVC)*, 2004.

[20] E. Tola, V. Lepetit, and P. Fua, "A fast local descriptor for dense matching," in *Conference on Computer Vision and Pattern Recognition*, Alaska, USA, 2008.