International Conference on Pattern Recognition, ICPR08,
Tampa Convention Center, Florida, USA, 2008.

A Novel 3-D Motion Estimation Approach to Virtual Viewpoint Control

Ying Kin YU , Siu Hang OR[1], Kin Hong WONG[1] and Kai Ki LEE[2]

[1]*Department of Computer Science & Engineering, The Chinese University of Hong Kong*
[2]*Department of Information Engineering, The Chinese University of Hong Kong*

ykyu.hk@gmail.com, { shor, khwong}@cse.cuhk.edu.hk, kklee6@ie.cuhk.edu.hk

Abstract

The novelty of this paper is the introduction of the Interacting Multiple Model Probabilistic Data Association Filter (IMMPDAF) to the pose tracking problem. The Interacting Multiple Model (IMM) technique allows the existence of more than one dynamic system in the filtering process and in return leads to improved accuracy and stability even under abrupt motion changes. The Probabilistic Data Association (PDA) framework makes the automatic selection of measurement sets possible, resulting in enhanced robustness to occlusions and moving objects. As the PDA associates stereo correspondences probabilistically, the explicit establishment of stereo matches is not necessary except during initialization, and the point features presence in the outer region of the stereo image pair can be utilized. The performance is demonstrated by applying the pose information to control cameras in a virtual environment.

1. Introduction

The research presented in this paper belongs to the category of Motion from Motion (MFM). The main concern is the recovery of camera position and orientation but not the 3-D structure. This is significantly different from traditional Structure from Motion (SFM) algorithms [4] [7] [13]. Despite the work that requires the existence and detection of a plane in the 3-D scene [6], the algorithm by Soatto *et. al.* [1] can be regarded as the seminal work of MFM according to our definition. In that, the essential constraint in epipolar geometry is applied to an EKF to directly compute the pose sequence from monocular images. As the consistency of the consecutive camera matrices can be guaranteed, the 3-D camera motion can be extracted afterwards. Yu *et. al.* proposed the EKF-based approaches that recover 3-D camera motion from monocular [2] and stereo images [5] using the trifocal tensor. As keeping track of the structural information is no longer required, putting these MFM algorithms into real usages is relatively easy and convenient.

A robust recursive MFM algorithm that recovers camera motion from a stereo image sequence based on the Interacting Multiple Model Probabilistic Data Association Filter (IMMPDAF) technique [11][12] is proposed in this article. The IMMPDAF computes the state estimates using multiple Probabilistic Data Association Filters (PDAFs), each of them describing a unique motion dynamic, and provides a probabilistic framework for the PDAFs to interact. The PDAF is able to account for the uncertainties of the measurement origins. Measurements acquired are checked against a validation region and the association probabilities of the validated measurements are computed, with which the final state is estimated. The IMMPDAF was originally designed to track a single target in a randomly distributed cluttered environment by the combination of multiple trajectory models with measurements from a radar and an infrared sensor [12]. From the literature we have encountered, it is believed we are the pioneers in incorporating the IMMPDAF framework into the latest model-free method for 3-D motion recovery.

The main advantages of the proposed approach are summarized as follow:

1. Robust operation even when one of the stereo cameras is partially blocked. As the measurements are associated across the views probabilistically, the explicit matching of stereo correspondences is no longer necessary except during initialization.

2. Considering multiple motion dynamics and abrupt motion changes. A number of hypotheses on the camera motion are enabled in our filter. The best motion model is "chosen" by the IMM algorithm in a probabilistic way at each filtering cycle. The highest accuracy on the recovered camera pose can thus be achieved due to the automatic application of the motion constraint.

2. Algorithm overview

An overview of the proposed pose tracking algorithm is shown in Fig. 1. A stereo camera set up is assumed and the Kanade-Lucas-Tomasi (KLT) tracker[9] is employed to extract feature points and track them in the succeeding images. The Interacting Multiple Model Probabilistic Data Association Filter (IMMPDAF)[11][12] is adopted to acquire the pose information from stereo image sequences. It consists of multiple Probabilistic Data Association Filters (PDAFs) [11] executing in parallel.

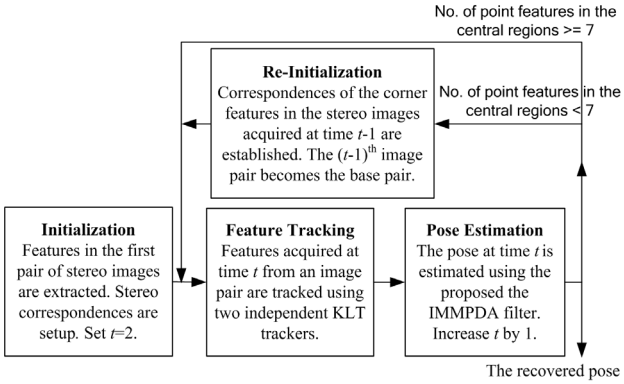

Fig.1. A flow chart giving an outline of the proposed algorithm.

To account for the uncertainty of the origin of the feature

coordinates, the PDAF associates probabilistically all the point features within the scene. At each time-step, a validation region is set up and the probability of each validated measurement for being correct is computed. The measurements remained are considered as outliers that arise either from the inaccuracy of the feature trackers or point features on a moving object in the static scene. The filter state is estimated based on the association probabilities and validated point features.

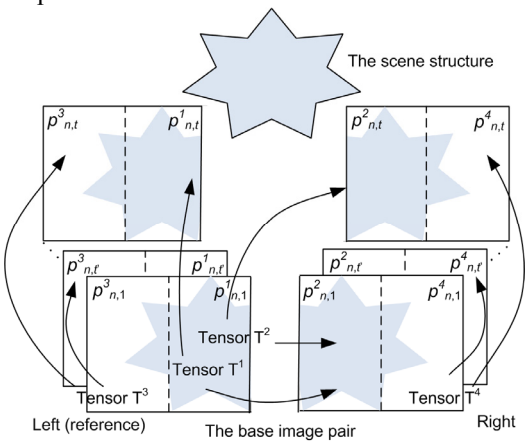

Fig.2 The arrangement of the stereo image pairs and partitioning of the views.

Fig. 2 illustrates the arrangement of image views for the PDAFs. A pair of images is divided into four parts. Two of them are the inner regions of the left of the right view, which are denoted by $p^1_{n,t}$ and $p^2_{n,t}$, respectively. They completely overlap with each other and matching of stereo correspondences is possible. The remaining parts are the outer regions of the stereo view, denoted by $p^3_{n,t}$ and $p^4_{n,t}$, that cover the non-overlapping portions of the stereo view. $p^1_{n,t}$, $p^2_{n,t}$, $p^3_{n,t}$ and $p^4_{n,t}$ compose of the 4 measurement sets in the PDAF. The PDAF "selects" the reliable sets of point features for filtering and associates the corresponding point features in the inner parts of the stereo images.

In order to recover the pose information directly without the explicit reconstruction of the scene structure, the trifocal tensor [2] is used.

As an improvement on the PDAF, the IMMPDAF computes the state estimates using multiple motion filters, each describing a unique motion dynamic and interacts with the others via a probabilistic framework. Three PDAFs are applied in our implementation. They are for *static* motion, and *mixed* motion and *planar* motion of constant velocity. Additional motion models can be incorporated depending on the actual application.

3. Implementation of the IMMPDAF

3.1 Dynamic system and measurement model

Let $\dot{x}_t, \dot{y}_t, \dot{z}_t, \dot{\alpha}_t, \dot{\beta}_t$ and $\dot{\gamma}_t$ be the translational velocities along the *x*, *y*, *z* axis and the angular velocities on the *x*, *y* and *z* axis, respectively. The state vector $\dot{\xi}_t(i)$ of the i^{th} motion filter (the i^{th} PDAF) is defined as

$$\dot{\xi}_t(i) = \begin{bmatrix} \dot{x}_t & \dot{y}_t & \dot{z}_t & \dot{\alpha}_t & \dot{\beta}_t & \dot{\gamma}_t \end{bmatrix}^T \quad (1)$$

With the assumption that sampling rate of the measurements is high, the dynamic system of the filter and the absolute pose M_t can be expressed using twist as

$$\dot{\xi}_t(i) = A(i)\dot{\xi}_{t-1}(i) + \eta_t \quad (2)$$

$$M_t = M_{t-1} e^{\tilde{\dot{\xi}}_t(i)} = M_{t-1}(I + \tilde{\dot{\xi}}_t(i)) \quad (3)$$

where $A(1) = I_{6X6}$, $A(2) = diag(\begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix})$ and $A(3) = 0_{6X6}$ are designed for the mixed motion (translation and rotation), planar motion (translation on the z-axis and rotation on the Pitch angle) and static motion, respectively. η_t is the zero-mean Gaussian noise with covariance Q_t. $\tilde{\dot{\xi}}_t(i)$ is the matrix form of $\dot{\xi}_t(i)$. The measurement equations of the filter are defined as

$$\varepsilon_t(k) = g_t(M_t, k) + \upsilon_t(k) \quad (4)$$

$$g_t(M_t, k) = \begin{bmatrix} u^k_{1,t} & v^k_{1,t} & \dots & u^k_{n,t} & v^k_{n,t} & \dots & u^k_{N,t} & v^k_{N,t} \end{bmatrix}^T \quad \text{for } 1 \le k \le 4 \in \mathrm{N} \quad (5)$$

$$[U^1_{n,t}]^c = [U^1_{n,1}]^a [l^2_{n,1}]_b [\mathrm{T}^1]^{bc}_a, \; [U^2_{n,t}]^c = [U^1_{n,1}]^a [l^2_{n,1}]_b [\mathrm{T}^2]^{bc}_a$$

$$[U^k_{n,t}]^c = [U^k_{n,1}]^a [l^k_{n,t'}]_b [\mathrm{T}^k]^{bc}_a \quad \text{for } 3 \le k \le 4 \in \mathrm{N} \quad (6)$$

$g_t(M_t, k)$ is the $N \times 1$ output function that transfers the coordinates of N point features belonging to the k^{th} measurement set from the base image pair to the t^{th} pair. It is actually the trifocal tensor point-point-point transfer function and has been given in (6), which is presented in the tensor notation. $\upsilon_t(k)$ represents the zero-mean Gaussian noise, having covariance $R_t(k)$, imposed on the images captured. T^k is the trifocal tensor that encapsulates the geometric relations among three view. $U^k_{n,t}$ is the normalized homogenous form of $p^k_{n,t}$ such that $U^k_{n,t} = \begin{bmatrix} \bar{u}^k_{n,t} & \bar{v}^k_{n,t} & \bar{w}^k_{n,t} \end{bmatrix}^T = \begin{bmatrix} u^k_{n,t}/f & v^k_{n,t}/f & 1 \end{bmatrix}^T$. The relation between tensor T^k and matrix M_t, and the construction of line $l^k_{n,t}$ can refer to [10].

3.2 The filtering equations

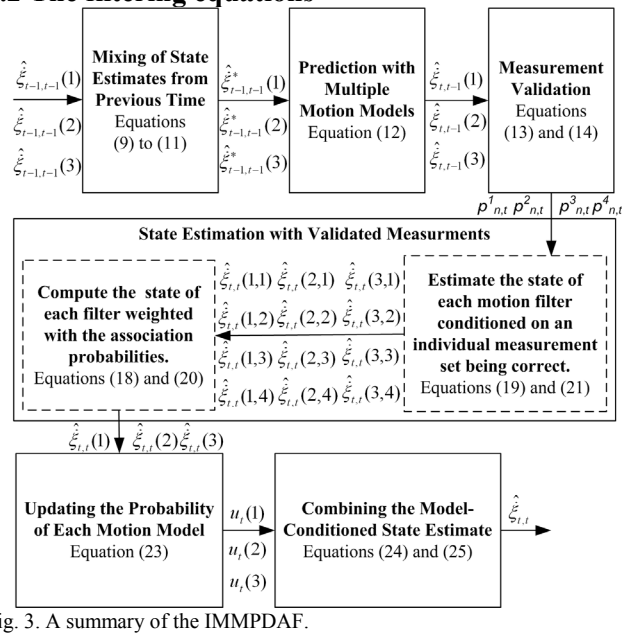

Fig. 3. A summary of the IMMPDAF.

At the beginning, estimates of different motion filters from the previous time-step $\hat{\hat{\xi}}_{t-1,t-1}(i)$, associated with covariance $P_{t-1,t-1}(i)$, are mixed according to the 3×3 switching matrix *J(i,j)* and the likelihood $u_{t-1}(i)$

$$\hat{\hat{\xi}}^{*}_{t-1,t-1}(i)=\frac{1}{u^{*}_{t}(i)}\sum_{j}J(i,j)u_{t-1}(j)\hat{\hat{\xi}}_{t-1,t-1}(j) \tag{7}$$

$$P^{*}_{t-1,t-1}(i)=\frac{1}{u^{*}_{t}(i)}\sum_{j}J(i,j)u_{t-1}(j)\Big(P_{t-1,t-1}(j)+ [\hat{\hat{\xi}}_{t-1,t-1}(j)-\hat{\hat{\xi}}_{t-1,t-1}(i)][\hat{\hat{\xi}}_{t-1,t-1}(j)-\hat{\hat{\xi}}_{t-1,t-1}(i)]^{T}\Big) \tag{8}$$

$$u^{*}_{t}(i)=\sum_{j}J(i,j)u_{t-1}(j) \tag{9}$$

Then the predicted state $\hat{\hat{\xi}}_{t,t-1}(i)$, having covariance $P_{t,t-1}(i)$, is computed

$$\hat{\hat{\xi}}_{t,t-1}(i)=A(i)\hat{\hat{\xi}}^{*}_{t-1,t-1}(i)$$
$$P_{t,t-1}(i)=A(i)P^{*}_{t-1,t-1}(i)A(i)^{T}+Q_{t} \tag{10}$$

After that, the measurements of feature set *k* predicted by the above models are combined using the predicted absolute pose $\hat{M}_{t,t-1}(i,j)$

$$\hat{\varepsilon}_{t,t-1}(k)=\sum_{i}\sum_{j}J(i,j)u_{t-1}(j)g_{t}(\hat{M}_{t,t-1}(i,j),k) \tag{11}$$

$\hat{\varepsilon}_{t,t-1}(k)$ represents the predicted coordinates after mixing. It is validated and thus should satisfy

$$[\varepsilon_{t}(k)-\hat{\varepsilon}_{t,t-1}(k)]^{T}\,\overline{S}_{t}(k)^{-1}[\varepsilon_{t}(k)-\hat{\varepsilon}_{t,t-1}(k)]<G^{2} \tag{12}$$

G is the standard deviation of the gate. The determination of $\left|\overline{S}_{t}(k)\right|$ can be found in [18]. Physically, the validation region is set to the largest volume among the three possible choices from the models. Each validated set of measurements has a corresponding association probability $B_{t}(k)$

$$B_{t}(k)=e_{t}(k)\left[b_{t}+\sum_{k}e_{t}(k)\right]^{-1}\quad B_{t}(0)=b_{t}\left[b_{t}+\sum_{k}e_{t}(k)\right]^{-1} \tag{13}$$

with
$$e_{t}(k)=(P_{G})^{-1}\mathrm{N}[r_{t}(k);0,S_{t}(k)] \tag{14}$$
$$b_{t}=L(1-P_{D}P_{G})(P_{D}P_{G}V_{t}(k))^{-1} \tag{15}$$

$B_{t}(0)$ is the probability that none of the measurement sets are correct. $\mathrm{N}[r_{t}(k);0,S_{t}(k)]$ is the normal probability density function. $r_{t}(k)$ is the measurement innovation associated with variance $S_{t}(k)$. $V_{t}(k)$ is the volume of the validation gate. P_{D} and P_{G} are respectively the probability for the scene point features being observed by the cameras and the probability for the features lying in the validation region. *L* is the number of valid measurement sets.

The measurements passed through the validation gate and the association probabilities $B_{t}(k)$ are used for state estimation

$$\hat{\hat{\xi}}_{t,t}(i)=\sum_{k}B_{t}(k)\hat{\hat{\xi}}_{t,t}(i,k) \tag{16}$$

$$\hat{\hat{\xi}}_{t,t}(i,k)=\hat{\hat{\xi}}_{t,t-1}(i)+W_{t}(i,k)r_{t}(i,k) \tag{17}$$

The corresponding covariances are computed by

$$P_{t,t}(i)=B_{t}(0)P_{t,t-1}(i)+\sum_{k}B_{t}(k)P_{t,t}(i,k)+ \sum_{k}B_{t}(k)\hat{\hat{\xi}}_{t,t}(i,k)\hat{\hat{\xi}}_{t,t}(i,k)^{T}-\hat{\hat{\xi}}_{t,t}(i)\hat{\hat{\xi}}_{t,t}(i)^{T} \tag{18}$$

$$P_{t,t}(i,k)=[I-W_{t}(i,k)\nabla g_{M}]P_{t,t-1}(i) \tag{19}$$

where $W_{t}(i,k)$ is the gain of the filter

$$W_{t}(i,k)=P_{t,t-1}(i,k)\nabla g_{M}{}^{T}\left[\nabla g_{M}P_{t,t-1}(i,k)\nabla g_{M}{}^{T}+R_{t}(k)\right]^{-1} \tag{20}$$

∇g_{M} is the Jacobian of the point transfer function $g_{t}(M_{t},k)$ evaluated at $\hat{\hat{\xi}}_{t,t-1}(i,k)$. Following the filtering step, the probability of each motion filter $u_{t}(i)$ is updated

$$u_{t}(i)=\kappa u^{*}_{t}(i)\Lambda_{t}(i) \tag{21}$$

κ is a normalization factor such that $\sum_{i}u_{t}(i)=1$. $\Lambda_{t}(i)$ is the joint probability density function of the innovations and its computation can refer to [18]. Lastly, the usable output state vector $\hat{\hat{\xi}}_{t,t}$ and covariance $P_{t,t}$ are generated

$$\hat{\hat{\xi}}_{t,t}=\sum_{i}u_{t}(i)\hat{\hat{\xi}}_{t,t}(i) \tag{22}$$

$$P_{t,t}=\sum_{i}u_{t}(i)\Big(P_{t,t}(i)+[\hat{\hat{\xi}}_{t,t}(i)-\hat{\hat{\xi}}_{t,t}][\hat{\hat{\xi}}_{t,t}(i)-\hat{\hat{\xi}}_{t,t}]^{T}\Big) \tag{23}$$

4. Experiments and Results

4.1 Synthetic data experiments

A large synthetic structure having 1000 randomly distributed feature points was generated. The stereo rig was moving in the structure and its motion was made up of 5 segments consisting of mixed (rotation and translation) and static motion. The parameters were random in the range from -1.0 to 1.0 degrees per frame for rotation and -0.025 to +0.025 meters per frame for translation. A 2-D zero-mean Gaussian noise of 0.5 pixel standard deviation was imposed. The length of each synthetic sequence was 100 frames. The moving path of the rig was long enough such that appearing and disappearing of feature points occurred naturally. To simulate presence of moving objects in the scene, groups of randomly moving point features were injected. The proposed IMMPDAF algorithm, the proposed PDAF approach (i.e. a variation of the IMMPDAF method that uses a single motion filter), the tensor-based EKF by Yu *et. al.* [5], and the traditional model-based EKF [8], in which the 3-D structure was assumed known, were implemented in Matlab and run on a Pentium IV 2GHz machine to estimate the camera motion. A total of 100 independent tests were carried out. Table I summarizes the performance of the algorithms.

4.2 Real image experiments

Real image sequences were used to test the proposed approach. An image sequence, which was 100-frame long, was captured using a hand-held stereo rig in a corner of the laboratory. The motion of the cameras was arbitrary and the degree of disturbance by foreign objects was severe. As no ground truth was available, only a visual check on the results could be made. Fig. 4 shows the real sequence and virtual

reality sequence with the recovered camera motion. Due to limited space, only the right view of the two stereo image sequences is included. One can notice that both the original and recovered motion were consistent with each other. More results can be found in the demonstration video http://www.cse.cuhk.edu.hk/~vision

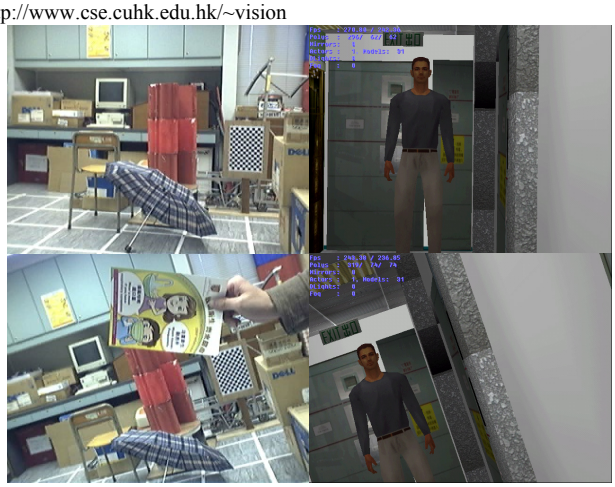

Fig. 4. Application of the motion recovered from the hand-held sequence to virtual reality. The 1st row shows the right view of the original (left) and resulting virtual (right) stereo sequence. The 2nd row illustrates the right view of the 77th image pair of the real (left) and virtual reality (right) sequence.

5. Conclusion

An innovative algorithm that acquires 3-D camera pose from a stereo image sequence based on the interacting multiple model probabilistic data association filter (IMMPDAF) has been described in this paper. Thanks to the probabilistic association of the point features across the stereo view, all corner features present in the images can be considered in the filtering process, no matter whether these features have or do not have stereo correspondences. The use of multiple motion filters allows motion constraints to be applied automatically, achieving the highest precision on the recovered camera pose and, at the same time, making the algorithm robust to abrupt motion changes. Our PDAF approach, which also outperformed an existing algorithm, can be regarded as a tradeoff between accuracy and computation efficiency. Real-time implementation of the proposed IMMPDAF and PDAF method is possible. The real image experiment shows that the IMMPDAF algorithm was accurate in the presence of moving objects and partial occlusion compared to the ground truth data. The proposed approach has a great potential to be used in a wide range of multimedia applications in addition to virtual reality.

6. Acknowledgement

This work was supported in part by Direct Grants under project codes 2050350 and 2050410 from the Faculty of Engineering, The Chinese University of Hong Kong.

7. References

[1] S.Soatto, R.Frezza and P.Perona, "Motion estimation on the essential manifold", presented at European Conf. Comput. Vision, Stockholm, Sweden, May 1994.

[2] Y.K.Yu, K.H.Wong, M.M.Y.Chang and S.H.Or, "Recursive camera motion estimation with the trifocal tensor", *IEEE Trans. Syst., Man, Cybern. B: Cybern.*, vol. 36, no. 5, pp. 1081- 1090, October 2006.

[3] T.J.Broida, S.Chandrashekhar and R.Chellappa, "Recursive 3-D motion estimation from a monocular image sequence", *IEEE Trans. Aerosp. Electron. Syst.*, vol. 26, no. 4, pp. 639-656, July 1990.

[4] A.J.Davison and D.W.Murray, "Simultaneous localization and map-building using active vision", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 7, pp. 865-880, July 2002.

[5] Y.K.Yu, K.H.Wong, S.H.Or and M.M.Y.Chang, "Recursive recovery of position and orientation from stereo image sequences without three-dimensional structures", *in Prof .IEEE Conf. Comput. Vision Pattern Recognit.*, vol. 1, pp. 1274-1279, New York, U.S.A., June 2006.

[6] M.Irani, B.Rousso and S.Peleg, "Recovery of ego-motion using image stabilization", in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, pp. 454-460, Seattle, USA, June 1994.

[7] D.Nister, "Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors", in *Proc. European Conf. Comput. Vision*, vol. 1, pp. 649-663, Ireland, June 2000.

[8] V.Lippiello, B.Siciliano and L.Villani, "Position and orientation estimation based on Kalman filtering of stereo images", in *Proc. IEEE Intl. Conf. Control Applications*, pp. 702-707, Mexico City, 2001.

[9] S.Baker, R.Gross, T.Ishikawa and I. Matthews, "Lucas-kanade 20 years on: A unifying framework: Part 2" Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-03-01, February 2003.

[10] R.Hartley and A.Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[11] Y.Bar-Shalom and T.E.Fortmann, *Tracking and data association*, Academic-Press, Boston, 1988.

[12] A.Houles and Y.Bar-Shalom, "Multisensor tracking of a maneuvering target in clutter", *IEEE Trans. Aerosp. Electron. Syst.*, vol. 25, no. 2, March 1989.

[13] A.J.Davison, I.D.Reid, N.D.Molton and O.Stasse, "MonoSLAM: Real-time single camera SLAM", *IEEE Trans. Pattern Anal. Machine Intell.*. , vol. 29, no. 6, pp. 1052-1067, June 2007.

TABLE I: A SUMMARY OF ALGORITHM PERFORMANCE

	The proposed IMMPDAF	The proposed PDAF	Yu's tensor-based EKF	Traditional Model-based EKF
Percentage of convergence	93.0%	83.0%	83.0%	98.0%
Average percentage of the accumulated total rotation errors (Diverged cases excluded)	4.3210%	4.2411%	6.5129%	1.7914%
Average percentage of the accumulated total translation errors (Diverged cases excluded)	9.0790%	8.0303%	13.6207%	3.5380%
Time required to process one point feature in an image	0.0051s	0.0009s	0.0013s	0.0010s

A table summarizing the performance of the algorithms under comparison in the synthetic experiment.