Controlling Virtual Cameras Based on a Robust Model-free Pose Acquisition Technique

Ying Kin Yu, Kin Hong Wong, Siu Hang Or and Junzhou Chen

***Abstract*—This paper presents a novel method that acquires camera position and orientation from a stereo image sequence without prior knowledge of the scene. To make the algorithm robust, the Interacting Multiple Model Probabilistic Data Association Filter (IMMPDAF) is introduced. 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. In addition to the IMMPDAF, the trifocal tensor is employed in the computation so that the step of reconstructing the 3-D models can be eliminated. This further guarantees the precision of estimation and computation efficiency. Real stereo image sequences have been used to test the proposed method in the experiment. The recovered 3-D motions are accurate in comparison with the ground truth data and have been applied to control cameras in a virtual environment.**

***Index Terms:* Pose tracking, Virtual reality, Augmented Reality, Interacting multiple model, Probabilistic data association, Trifocal tensor, Stereo vision, Multimedia processing**

I. Introduction

Virtual camera control can be understood as moving the user's point of view in a 3-D virtual environment [1]. This task is also known as 3-D scene exploration or viewpoint placement in the computer graphics community. Controlling the camera viewpoint is crucial for a wide range of applications, for instances, movie making for entertainment and design of camera motions in animations. Given the 3-D structures, the problem of camera motion recovery can be solved using the model-based approaches [2][3][7][31], which are well-known and have good performance under a controlled environment. If prior information on the scene is not available, traditional Structure from Motion (SFM) algorithms [4][5], which simultaneously estimate the scene structure and pose information, are required. Lee and Kay [8] used an EKF to estimate the pose as well the structure of an object with a stereo camera system. The series of methods in [4][5][9][10][11][12][13] recover both the structure and motion simultaneously using Kalman filters.

The research presented in this paper belongs to a different category: Motion from Motion (MFM), as mentioned in [16], in which the main concern is the camera position and orientation but not the 3-D structure. To be more precise, MFM algorithms have the capability of estimating 3-D camera motion directly from 2-D image motion without the explicit reconstruction of the scene structure, even though the 3-D model structure is not known in prior [14][15][17][18]. As keeping track of the structural information is no longer required, putting these types of algorithms into real applications is relatively easy and convenient.

A robust recursive MFM algorithm that recovers camera motion from a stereo image sequence for virtual reality based on the Interacting Multiple Model Probabilistic Data Association Filter (IMMPDAF) technique [21][22] 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 uncertainty of the measurement origin. 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 [22]. Recently, such a concept has been adopted to the extraction of cavity contours from ultra sound images [24]. From the literature we have encountered, it is believed we are the first to incorporate the IMMPDAF framework into the latest model-less method [6][15] for 3-D motion recovery. The proposed approach is able to achieve high accuracy and stability under occlusions, moving objects, and the presence of abrupt motion changes. Compared to our previous SFM-based method that employs only the Interacting Multiple Model (IMM) [13], its performance has been greatly improved due to the probabilistic association of point features and model-free nature of the algorithm. Strengths of our algorithm are:

Robust operation even when one of the stereo cameras is partially blocked. By making use of the Probabilistic Data Association (PDA) method [21], the proposed algorithm is able to take all available corner features into account in the filtering process elegantly no matter whether these features do or do not have stereo correspondences. Also, the PDA formulation enables our approach to choose reliable measurements, reject outliers and give weighting factor to the selected set of corner features, leading to an increase in robustness of the proposed method.

Considering multiple motion dynamics and abrupt motion changes. With the Interacting Multiple Model (IMM) algorithm [22], a number of hypotheses on camera motion are enabled in our filter. The best motion model is "chosen" by the IMM algorithm in a probabilistic way. The highest accuracy on the recovered camera pose can thus be achieved due to the automatic application of the motion constraint. Moreover, the mechanism of "switching" among different models allows the presence of abrupt motion changes. In specific applications, the stability can be improved compared to the EKF-based approaches for SFM that use only one motion model.

Recovering position and orientation without the explicit reconstruction of 3-D structure. Our novel method is able to compute the pose information directly from a stereo image

sequence without the step of reconstructing the 3-D model. To achieve the goal, the trifocal tensor point transfer function is applied to the measurement model of the filter. At each filtering cycle, only the 6 parameters of the pose, instead of N + 6 parameters of both the structure and motion, are required to be estimated.

This paper is organized as follows. The pose acquisition problem is defined in Section II. An overview of the IMMPDAF pose tracking algorithm is then illustrated in Section III. The details of the application of trifocal tensor, the formulation of the PDAF and IMMPDAF are given in Section IV. In Section V, the comparison among our approaches, the trifocal tensor-based EKF by Yu *et. al.* [15] and the standard model-based EKF [3] using synthetic data is presented. Also, experimental results of our IMMPDAF approach with a real stereo image sequence having ground truth are shown. Application of our method to virtual reality is illustrated before making a conclusion.

II. Modeling of the Pose Tracking Problem

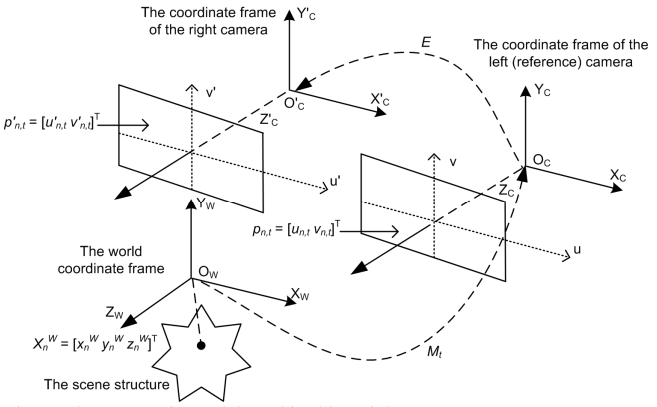

Fig. 1. The geometric model used in this article.

The geometric setup of our stereo system is shown in Fig. 1. Let $X_n^W = [x_n^W \quad y_n^W \quad z_n^W]^T$ be the coordinates of the n^{th} model point in the world coordinate frame. Therefore

$$\begin{bmatrix} \tilde{u}_{n,t} \\ \tilde{v}_{n,t} \\ \tilde{w}_{n,t} \end{bmatrix} = K[I_{3\times3} \quad 0_{3\times1}]M_t \begin{bmatrix} x_n^W \\ y_n^W \\ z_n^W \\ 1 \end{bmatrix}, \begin{bmatrix} \tilde{u}'_{n,t} \\ \tilde{v}'_{n,t} \\ \tilde{w}'_{n,t} \end{bmatrix} = KEM_t \begin{bmatrix} x_n^W \\ y_n^W \\ z_n^W \\ 1 \end{bmatrix} \quad (1)$$

where E is a 3×4 matrix representing the rigid transformation between the two cameras. K is a 3×3 matrix that encodes the intrinsic parameters of a camera such as the focal length f. M_t transforms the 3-D structure from the world frame to the reference camera at time instance t. It can be parameterized into x_t, y_t, z_t, $\alpha_t, \beta_t, \gamma_t$, which are respectively the translations in the x, y and z direction and the rotations about the x, y and z axis, using the twist representation [27]. The actual image coordinates $p_{n,t} = [u_{n,t}, v_{n,t}]^T$ on the left view and $p'_{n,t} = [u'_{n,t}, v'_{n,t}]^T$ on the right view are respectively given by

$$\begin{bmatrix} u_{n,t} \\ v_{n,t} \end{bmatrix} = \begin{bmatrix} \tilde{u}_{n,t} / \tilde{w}_{n,t} \\ \tilde{v}_{n,t} / \tilde{w}_{n,t} \end{bmatrix}, \quad \begin{bmatrix} u'_{n,t} \\ v'_{n,t} \end{bmatrix} = \begin{bmatrix} \tilde{u}'_{n,t} / \tilde{w}'_{n,t} \\ \tilde{v}'_{n,t} / \tilde{w}'_{n,t} \end{bmatrix} \quad (2)$$

The objective of the proposed algorithm is to compute the 3-D camera motion, i.e. M_t, at each time-step recursively given only the image measurements $p_{n,t}$ and $p'_{n,t}$.

III. Outline of the Algorithm

A. Initialization

An overview of the proposed pose tracking algorithm is shown in Fig. 2. The Kanade-Lucas-Tomasi (KLT) tracker [20] is employed to extract feature points and track them in the succeeding images. Starting from the 1st image pair at time t=1, and stereo images are matched with each other to establish the stereo correspondences. To perform stereo matching, features from the left and right images are used to find the fundamental matrix (F) [26] and the Random Sample Consensus (RANSAC) robust estimator [23].

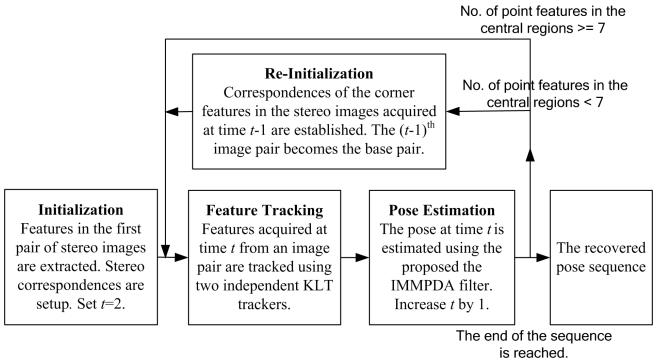

Fig.2. A flowchart giving an outline of the proposed algorithm.

A guided search is then performed. The pair of points, say $p_{n,t}$ and $p'_{n,t}$, is regarded as matched if the distance between $p_{n,t}$ and the epipolar line of $p'_{n,t}$ is the shortest and has the highest correlation value. The set of newly acquired matches is used to improve the accuracy of F, which can in turn be applied to the next guided search. The process is repeated until no more point matches can be found [25]. With known camera intrinsic parameters K, the extrinsic parameters E of the stereo system can be extracted from F according to [26].

The values from F are then used as part of the initial guess of tensors T^1 and T^2. A portion of the erroneous point features from the KLT tracker are rejected during the setup of stereo correspondences in the base image pair, as they are unable to have putative matches across the left and right views.

B. Pose acquisition

The Interacting Multiple Model Probabilistic Data Association Filter (IMMPDAF) [21][22] is adopted to acquire the pose information from stereo image sequences. It is an extension of the Probabilistic Data Association Filter (PDAF) [21], a suboptimal Bayesian algorithm assuming that there is only one target of interest in the measurements. 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. As an improvement on the PDAF, the IMMPDAF computes the state estimates using multiple motion filters. Each filter, basically a PDAF, describes a unique motion dynamic and interacts with the others via a probabilistic framework.

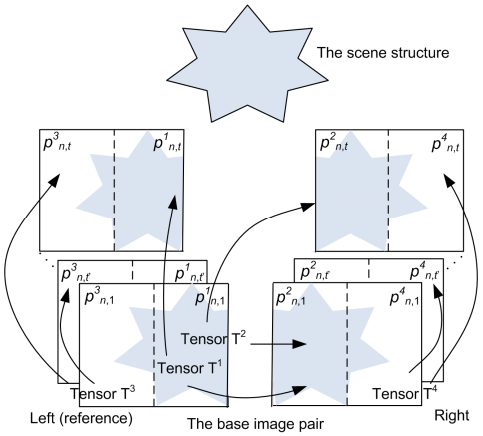

Fig.3 Stereo image pair arrangements and partitioning of the views.

Fig. 3 illustrates the arrangement of image views in our pose tracking algorithm. The partition is based on the property that point features observed from a stereo camera can or cannot have stereo correspondences. As a stereo camera captures two images at a time, a pair of images is divided into four parts. Two of them are the inner regions of the left and 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 and the setup of stereo matches is impossible. $p^1_{n,t}$, $p^2_{n,t}$, $p^3_{n,t}$ and $p^4_{n,t}$ compose of the 4 measurement sets in the PDAF. The major role of the PDAF is to provide a mechanism to "select" the reliable sets of point features for filtering. An auxiliary function of PDAF is to associate 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 [26] is required. The 4 sets of measurements in a pair of stereo views are linked together by a total of 4 trifocal tensors as follows. The first tensor T^1 establishes the geometric relation among the inner left $p^1_{n,1}$ and right view $p^2_{n,1}$ of the base pair, and the inner left view $p^1_{n,t}$ of the current stereo image. The second tensor T^2 sets up a relation among the inner left $p^1_{n,1}$ and right view $p^2_{n,1}$ of the base pair, and the inner right view $p^2_{n,t}$ of the current images. The third T^3 and the fourth T^4 tensor connect the outer left views $p^3_{n,1}$, $p^3_{n,t'}$, $p^3_{n,t}$ and outer right views $p^4_{n,1}$, $p^4_{n,t'}$, $p^4_{n,t}$, respectively. t' is an integer such that $1<t'<t$ and is chosen as 10 in the experiment. These 4 tensors are incorporated into the measurement model of the PDAF.

The choice of t' determines how far the first two image views in the outer partitions constrained by tensors T^3 and T^4 are separated. A good choice of t' could improve both the accuracy and robustness of the proposed approach. If t' is set to a small value, say 2, the separation among image views $p^3_{n,1}$, $p^3_{n,t'}$, $p^3_{n,t}$ and among $p^4_{n,1}$, $p^4_{n,t'}$, $p^4_{n,t}$ may be too small when the motion of the cameras is very slow, resulting in a near degenerate condition of tensors T^3 and T^4. Once happened, the proposed IMMPDAF can automatically be switched to make use of the measurements from the inner partitions and continue the operation. A better choice of t' could avoid these two tensors becoming degenerate so that point features in all four partitions can be fully utilized, enhancing the accuracy and stability of the algorithm.

In our implementation, three PDAFs are applied in parallel under the Interacting Multiple Model (IMM) framework. These filters are for static motion, and mixed motion and planar motion of constant velocity Additional motion models can be incorporated depending on the actual application of the proposed algorithm.

C. *Re-initialization*

While the cameras are in motion, the set of observable feature points is changing as new scene structures may appear and old ones may become out of sight. The coordinates of the corner features are input to the filter as measurements. The proposed filter is required to be bootstrapped once the number of available point features related by the trifocal tensors T^1 or T^2 in the central region of the stereo view is below 7. The reason is that a trifocal tensor is unable to be established with 6 or less point correspondences across 3 views. With 7 or more point correspondences, the trifocal constraint is able to characterize the rigid motion of the cameras. Under the stereo configuration, a degenerate situation of all the 4 trifocal tensors will only occur when the cameras are observing a pure planar surface, which is unlikely to happen in the reality. Such a situation will not cause the system to bootstrap but may lead the algorithm to diverge if the cameras do not move by following the original motion dynamic.

During re-initialization, the views at the current time-step are set as the new base image pair and the KLT tracker is restarted. The number of points constrained by the 4 trifocal tensors can be increased by shifting the base image pair forward. Matching of stereo correspondences is performed on the base pair as mention in Section III-A. Note that the fundamental matrix F is not necessary to be re-computed as the relative pose of the two cameras is assumed fixed.

IV. Structure-less 3-D Motion Recovery with IMMPDAF

A. The 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

$$\dot{\xi}_t(i) = \begin{bmatrix} \dot{x}_t & \dot{y}_t & \dot{z}_t & \dot{\alpha}_t & \dot{\beta}_t & \dot{\gamma}_t \end{bmatrix}^T \quad (3)$$

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

$$\dot{\xi}_t(i) = A(i)\dot{\xi}_{t-1}(i) + \eta_t \quad (4)$$

$$M_t = M_{t-1} e^{\tilde{\dot{\xi}}_t(i)} = M_{t-1}(I + \tilde{\dot{\xi}}_t(i)) \quad (5)$$

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 $\dot{\xi}_t(i)$ [27]. The measurement equations of the filter are defined as

$$\varepsilon_t(k) = g_t(M_t, k) + \upsilon_t(k) \quad (6)$$

$$g_t(M_t, k) = \begin{bmatrix} u_{1,t}^k & v_{1,t}^k & \dots & u_{n,t}^k & v_{n,t}^k & \dots & u_{N,t}^k & v_{N,t}^k \end{bmatrix}^T$$

$$\text{for } 1 \le k \le 4 \in \mathrm{N} \quad (7)$$

$$[U_{n,t}^1]^c = [U_{n,1}^1]^a [l_{n,1}^2]_b [\mathrm{T}^1]_a^{bc},\ [U_{n,t}^2]^c = [U_{n,1}^1]^a [l_{n,1}^2]_b [\mathrm{T}^2]_a^{bc}$$

$$[U_{n,t}^k]^c = [U_{n,1}^k]^a [l_{n,t'}^k]_b [\mathrm{T}^k]_a^{bc} \text{ for } 3 \le k \le 4 \in \mathrm{N} \quad (8)$$

$g_t(M_t, k)$ 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 transfer function and has been given in (8), 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 views [26] and is defined in Section III-B. Three corresponding points across the views related by tensor T^k form a relation known as point-point-point correspondence. $U_{n,t}^k$ is the normalized homogenous form of $p_{n,t}^k$ such that $U_{n,t}^k = \begin{bmatrix} \bar{u}_{n,t}^k & \bar{v}_{n,t}^k & \bar{w}_{n,t}^k \end{bmatrix}^T = \begin{bmatrix} u_{n,t}^k / f & v_{n,t}^k / f & 1 \end{bmatrix}^T$. The relation between tensor T^k and matrix M_t, and the construction of line $l_{n,t}^k$ can refer to [26].

B. The interacting multiple model probabilistic data association filter (IMMPDAF)

With the dynamic system and measurement model, the IMMPDAF can be implemented. A glance on the steps involved is demonstrated in Fig. 4. At the beginning, estimates of different motion filters from the previous time-step $\hat{\dot{\xi}}_{t-1,t-1}(i)$, associated with covariance $P_{t-1,t-1}(i)$, are mixed according to the 3×3 switching matrix $J(i,j)$ and the likelihood $u_{t-1}(i)$

$$\hat{\dot{\xi}}_{t-1,t-1}^*(i) = \frac{1}{u_t^*(i)} \sum_j J(i,j) u_{t-1}(j) \hat{\dot{\xi}}_{t-1,t-1}(j) \quad (9)$$

$$P_{t-1,t-1}^*(i) = \frac{1}{u_t^*(i)} \sum_j J(i,j) u_{t-1}(j) \Big(P_{t-1,t-1}(j) + [\hat{\dot{\xi}}_{t-1,t-1}(j) - \hat{\dot{\xi}}_{t-1,t-1}(i)][\hat{\dot{\xi}}_{t-1,t-1}(j) - \hat{\dot{\xi}}_{t-1,t-1}(i)]^T \Big) \quad (10)$$

$$u_t^*(i) = \sum_j J(i,j) u_{t-1}(j) \quad (11)$$

where $J(i,j)$ is the Markov model-switching probability that gives the jump probability from motion filter i to motion filter j. Then the predicted state $\hat{\dot{\xi}}_{t,t-1}(i)$, having covariance $P_{t,t-1}(i)$, is computed

$$\hat{\dot{\xi}}_{t,t-1}(i) = A(i)\hat{\dot{\xi}}_{t-1,t-1}^*(i) \quad (12)$$

$$P_{t,t-1}(i) = A(i)P_{t-1,t-1}^*(i)A(i)^T + Q_t$$

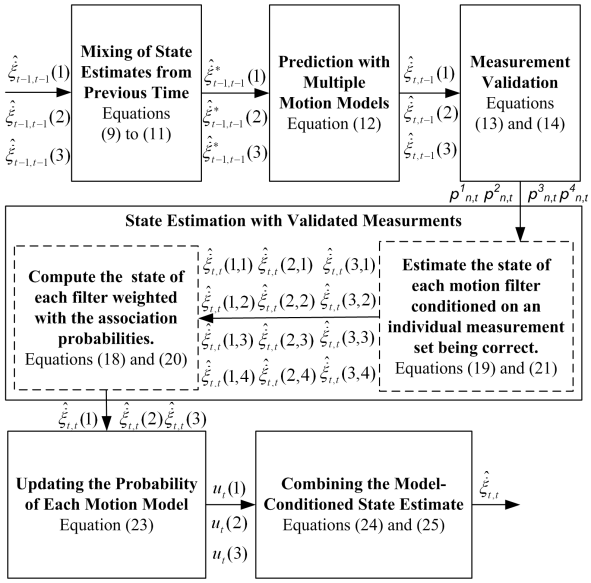

Fig. 4. A summary of the IMMPDAF method.

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

$$\hat{\varepsilon}_{t,t-1}(k) = \sum_i \sum_j J(i,j) u_{t-1}(j) g_t(\hat{M}_{t,t-1}(i,j), k) \quad (13)$$

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

$$[\varepsilon_t(k) - \hat{\varepsilon}_{t,t-1}(k)]^T \bar{S}_t(k)^{-1} [\varepsilon_t(k) - \hat{\varepsilon}_{t,t-1}(k)] < G^2 \quad (14)$$

G is the standard deviation of the gate. The determination of $|\bar{S}_t(k)|$ can be found in [22]. Physically, the validation region is set to the largest volume among the three possible choices from the models. Point features in the outer partitions and matched point pairs in the inner partitions are validated by the g-sigma gate as described by formula (14). The volume of the gate depends on the residual covariance of the measurements obtained. If the predicted 2-D positions of the

point features lie outside the area determined by the validation gate, those features will be regarded as outliers and the associated partitions will not be used for the correction of prediction in the filter.

Each validated set of measurements has a corresponding association probability $B_t(k)$

$$B_t(k) = e_t(k)\left[b_t + \sum_k e_t(k)\right]^{-1} \quad B_t(0) = b_t\left[b_t + \sum_k e_t(k)\right]^{-1} \tag{15}$$

with
$$e_t(k) = (P_G)^{-1}\mathrm{N}[r_t(k);0,S_t(k)] \tag{16}$$

$$b_t = L(1-P_D P_G)(P_D P_G V_t(k))^{-1} \tag{17}$$

$B_t(0)$ 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

$$\hat{\xi}_{t,t}(i) = \sum_k B_t(k)\hat{\xi}_{t,t}(i,k) \tag{18}$$

$$\hat{\xi}_{t,t}(i,k) = \hat{\xi}_{t,t-1}(i) + W_t(i,k)r_t(i,k) \tag{19}$$

The corresponding covariances are computed by

$$P_{t,t}(i) = B_t(0)P_{t,t-1}(i) + \sum_k B_t(k)P_{t,t}(i,k) + \sum_k B_t(k)\hat{\xi}_{t,t}(i,k)\hat{\xi}_{t,t}(i,k)^T - \hat{\xi}_{t,t}(i)\hat{\xi}_{t,t}(i)^T \tag{20}$$

$$P_{t,t}(i,k) = [I - W_t(i,k)\nabla g_M]P_{t,t-1}(i) \tag{21}$$

where $W_t(i,k)$ is the gain of the filter

$$W_t(i,k) = P_{t,t-1}(i,k)\nabla g_M{}^T\left[\nabla g_M P_{t,t-1}(i,k)\nabla g_M{}^T + R_t(k)\right]^{-1} \tag{22}$$

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

$$u_t(i) = \kappa u_t^*(i)\Lambda_t(i) \tag{23}$$

κ is a normalization factor such that $\sum_i u_t(i) = 1 \cdot \Lambda_t(i)$ is the joint probability density function of the innovations and its computation can refer to [22]. Lastly, the usable output state vector $\hat{\xi}_{t,t}$ and covariance $P_{t,t}$ are generated

$$\hat{\xi}_{t,t} = \sum_i u_t(i)\hat{\xi}_{t,t}(i) \tag{24}$$

$$P_{t,t} = \sum_i u_t(i)\left(P_{t,t}(i) + [\hat{\xi}_{t,t}(i) - \hat{\xi}_{t,t}][\hat{\xi}_{t,t}(i) - \hat{\xi}_{t,t}]^T\right) \tag{25}$$

The final state estimates rely more on the less noisy measurement sets.

V. Experiments and Results

A. *Experiments with synthetic data*

A 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. A 2-D zero-mean Gaussian noise of 0.5 pixel standard deviation was imposed. Projections of random point features were present on both the inner and outer partitions of the image planes. All partitions of the stereo views were filled with feature points unless the cameras moved out of or near to the boundaries of the 3-D structure. The moving path of the rig was long enough such that appearing and disappearing of feature points occurred naturally. The actual number of point features that could be observed from the stereo views depended on the position and orientation of the cameras. To simulate presence of moving objects in the scene, groups of randomly moving point features were injected. In addition, outliers were inserted into the synthetic sequences.

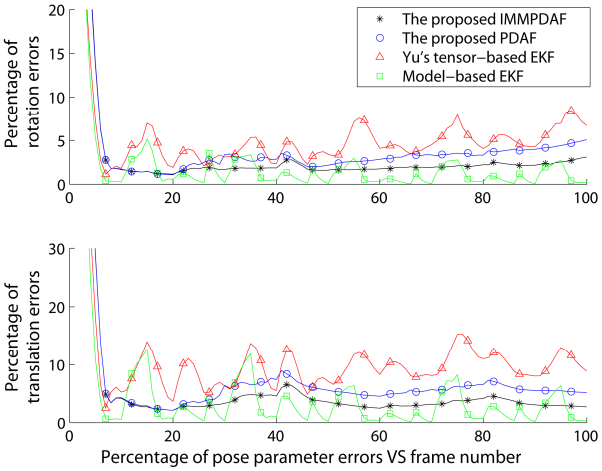

Fig. 5. The average percentages of accumulated rotation errors (top) and translation errors (bottom) against frame number of the algorithms under comparison. The diverged cases were excluded when computing the average values.

We are interested in the performance of the proposed algorithm in comparison with other methods that do not require the computation of 3-D structure in the pose acquisition process. 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.* [15] and the traditional model-based EKF [3], 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.

There are also other approaches [28] that we would like to compare with. Due to the dissimilarities in operation as well as implementation consideration, these algorithms are not included in the empirical comparison.

To test our approach with a more general setting, the planar motion filter, which is designed for robot motion, was disabled in this experiment so that the IMMPDAF only consisted of two motion filters. A total of 100 independent

tests were carried out.

In Fig. 5. The lines with asterisk (*), circle (○), triangle (△) and square (□) markers represent the proposed IMMPDAF approach, the PDAF method, the tensor-based EKF by Yu *et. al.* [15] and the traditional model-based EKF [3], respectively. The diverged cases were removed when plotting these graphs so as to make the average values meaningful and reasonable. Here the accumulated rotation error is defined as the difference between the actual and the recovered angles in the axis-angle representation while the accumulated translation error equals to the absolute difference between the actual and the recovered translation. One can observe that both of our IMMPDAF and PDAF approach were more accurate than the tensor-based EKF [15] that does not have the ability to switch among partitions probabilistically. They were able to achieve an error level comparable to the model-based method within the first 40 image frames. As the observable set of point features was changing in the synthetic environment, the pose acquired by our approaches might drift a bit, leading to a higher error compared to the model-based EKF.

TABLE I
COMPARISON OF ALGORITHMS IN THE SYNTHETIC EXPERIMENT.

*=Diverged cases excluded	Our IMM-PDAF	Our PDAF	Yu's tensor-based EKF	Model-based EKF
% of convergence	98.0	90.0	86.0	99.0
Average % of accumulated total rotation errors*	2.95	3.89	5.02	1.87
Average % of accumulated total translation errors *	5.53	7.23	10.44	4.24
Process 1 point feature in 1 image	0.0032s	0.0014s	0.0017s	0.0011s

Table I summarizes the performance of the algorithms in the experiment. The differences in performance among them are quite clear after the injection of outlying point features. Our IMMPDAF method was the most stable (from the percentage of convergence) except for the model-based EKF. It was even better than the PDAF approach since the synthetic motion contained motion discontinuities and the use of multiple motion models made itself able to resolve the case. The tensor-based method [15], on the other hand, is susceptible to the attacks of outliers and moving objects.

The computation time per point feature is also demonstrated. Since the IMMPDAF algorithm consisted of more than one PDAFs plus computation overhead, so it ran longer than our PDAF method. If a single PDAF was applied, its computation efficiency was higher than that of the tensor-based EKF [15] because the input measurement set was broken down into smaller groups. Compared to the classic model-based EKF, the speed of the PDAF algorithm, a MFM method, was just a bit slower.

Although the PDAF algorithm (a tradeoff between speed and accuracy) outperformed the tensor-based EKF, it was less precise and robust than the IMMPDAF. As the occurrences of independently moving point features was frequent and the change of motion was quite drastic, the benchmark model-based EKF did diverge in a few test cases.

B. *Experiments with real images*

The proposed IMMPDAF, together with the PDAF approach (a variation of the IMMPDAF method with a single motion filter), were tested for their robustness using a real stereo image sequence. The sequence is consisted of 190 frames and the robot was programmed to follow a path in the laboratory with disturbing moving objects.

Fig. 6. In the first row, images 1,2 (from left) is the input image pair, and 3,4, are generated synthetic results of the 1st image pair. The second row shows the input and result of the corresponding 160th image pair. See http://www.cse.cuhk.edu.hk/~vision/

Figs. 6 to 9 are the results of the first test sequence. A virtual reality video was successfully created using the camera motion extracted by our IMMPDAF. From Fig. 6 and the demonstration video, both the original and recovered motion were consistent with each other. When compared to the ground truth, the proposed methods were precise and the IMMPDAF algorithm could recover a less noisy pose sequence than the PDAF approach as indicated in Fig. 7. There is a sudden increase in translation error along the x-axis of the proposed IMMPDAF approach. This is due to the fact that moving objects were present in the scene. Once these moving objects were "detected", their effects were eliminated. One can noticed that the recovered translation x_t in Fig. 7 gradually returned to the correct value. The reaction time of the IMMPDAF to the moving objects can be tuned to obtain an optimal result.

The transitions of the motion filters are revealed in Fig. 8. In this test case, sudden stops were intentionally made in the robot motion. Our IMMPDAF algorithm switched to the static motion filter (SMF) at the 40th, 106th and 144th frame successfully. For the rest of the video sequence, the planar motion was selected most of the time due to the constrained robot movement. The use of the mixed motion filter (MMF) between the 88th and 124th frame was for the correction on the accumulated pose errors other than the Pitch angle rotation (β_t) and Z translation (z_t). Fig. 9 illustrates the probabilities associated with the measurement sets in the frequently selected planar motion filter. As moving objects were present in the scene, the likelihoods of some sets of measurements were significantly lowered from the 50th to 100th frame, and around the 125th frame. The IMMPDAF handled occlusions and moving objects successfully. In addition, we verified the recovered pose by inserting a virtual object (a virtual human) into the original video. As can be seen, the estimated motion fit nicely to create the illusion that the human is standing on

the red wing of a toy plane all the time. The augment reality video can be found at http://www.cse.cuhk.edu.hk/~vision/

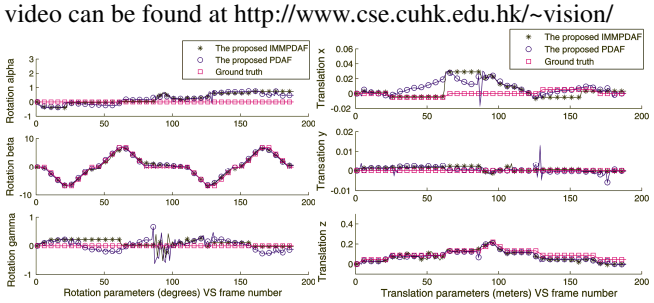

Fig. 7. The pose recovered with our IMMPDAF and PDAF approach. The results are compared with the ground truth values.

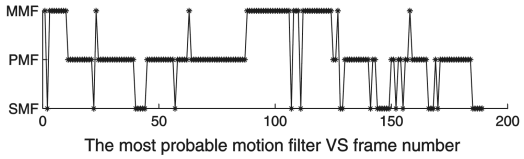

Fig. 8. The relation between the most probable motion filter and frame number in the IMMPDAF. MMF, PMF and SMF are the short forms of the mixed motion filter, planar motion filter and static motion filter, respectively.

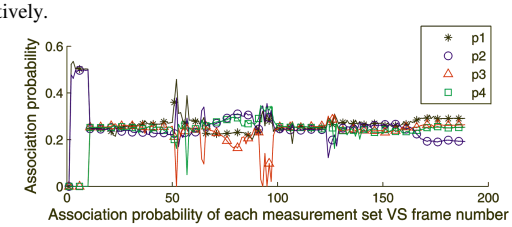

Fig. 9. The association probabilities of each measurement set of the planar motion filter computed by the IMMPDAF.

VI. Conclusion

An innovative method that acquires 3-D camera pose for virtual reality has been described in this paper. In the algorithm, the Interacting Multiple Model Probabilistic Data association Filter (IMMPDAF) is introduced to recover pose information from a stereo image sequence. 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 explicit searching of stereo matches is no longer necessary except during initialization. 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. The trifocal tensor embedded within the IMMPDAF enables the direct computation of the pose information by skipping the step of reconstructing the 3-D structure, even if the model of the scene is not available. The computation of our approach is thus optimized and its implementation becomes simple. Experimental results reveal that the stability of the proposed IMMPDAF method was comparable to the traditional model-based pose estimation algorithm, which requires known scene structure for calculation, and was much better than the latest Motion from Motion (MFM)-based extended Kalman filter (EKF) by Yu *et. al.* [15]. On the other hand, our Probabilistic Data Association Filter (PDAF) method that computes pose information with a single motion model can be regarded as a tradeoff between speed and accuracy. The real image experiment shows that both the IMMPDAF and PDAF algorithm were accurate in the presence of moving objects compared to the ground truth data. The estimated pose has successfully been applied to drive a pair of cameras in a virtual environment. The proposed approach has a great potential to be used in a wide range of multimedia applications such as the creation of augmented reality videos [29][30] in addition to virtual reality.

References

[1] C.Ware and S.Osborn, "Exploration and virtual camera control in virtual three dimensional environments", in *Proc. of the Symposium on Interactive 3-D Graphics*, Snowbird, Utah, pp. 175-183, 1990.

[2] D.G.Lowe, "Fitting parameterized three-dimensional models to images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 441-450, May 1991.

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

[4] A.Chiuso, P.Favaro, H.Jin and S.Soatto, "Structure from motion causally integrated over time", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 523-535, April 2002.

[5] A.Azarbayejani and A.P.Pentland, "Recursive estimation of motion, structure, and focal length", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, pp. 562-575, June 1995.

[6] Y.K.Yu, K.H.Wong, S.H.Or, and M.M.Y.Chang, "Robust 3-D Motion Tracking From Stereo Images: A Model-Less Method", IEEE Transactions on Instrumentation and Measurement, vol. 57, no. 3, pp. 622-630, March 2008

[7] Y.K.Yu, K.H.Wong and M.MY.Chang, "Pose Estimation for Augmented Reality Applications Using Genetic Algorithm", *IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics*, vol. 35, no. 6, pp. 1295- 1301, December 2005.

[8] S.Lee and Y.Kay, "An accurate estimation of 3-D position and orientation of a moving object for robot stereo vision: Kalman filter approach", in *Proc. of IEEE International Conference on Robotics and Automation*, pp. 414-419, Ohio, May 1990.

[9] T.J.Broida, S.Chandrashekhar and R.Chellappa, "Recursive 3-D motion estimation from a monocular image sequence", *IEEE Transactions on Aerospace and Electronic Systems*,vol. 26, no. 4, pp. 639-656,July 90.

[10] A.J.Davison and D.W.Murray, "Simultaneous localization and map-building using active vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 865-880, July 2002.

[11] J.Weng, N.Ahuja and T.S.Huang, "Optimal motion and structure estimation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 864-884, September 1993.

[12] Y.K.Yu, K.H.Wong and M.M.Y.Chang, "Recursive three-dimensional model reconstruction based on Kalman filtering", *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 35, no. 3, pp. 587-592, June 2005.

[13] Y.K.Yu, K.H.Wong and M.Y.Y.Chang, "Merging artificial objects with marker-less video sequences based on the interacting multiple model method", *IEEE Transactions on Multimedia,* vol. 8 no. 3, pp. 521-528, June 2006.

[14] Y.K.Yu, K.H.Wong, M.M.Y.Chang and S.H.Or, "Recursive camera motion estimation with the trifocal tensor", *IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics*, vol. 36, no. 5, pp. 1081- 1090, October 2006.

[15] 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. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 1274-1279, New York, U.S.A., June 2006.

[16] A.Chiuso and S.Soatto, "Mfm: 3-D Motion from 2-D motion causally integrated over time part I: Theory", presented at European Conference on Computer Vision, Bublin, June 2000.

[17] S.Avidan and A.Shashua, "Threading fundamental matrices", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 73-77, January 2001.

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

[19] A.J.Davison and N.Kita, "3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain", in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp.384-391, Kauai, December 2001.

[20] C.Tomasi and T.Kanade, "Detection and tracking of point features", Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

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

[22] A.Houles and Y.Bar-Shalom, "Multisensor tracking of a maneuvering target in clutter", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 25, no. 2, March 1989.

[23] M.A.Fischler and R.C.Bolles, "Random sample concensus: A paradigm for model fitting with applications to image analysis and automated cartography", *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, June 1981.

[24] P.Abolmaesumi and M.R.Sirouspour, "An interacting multiple model probabilistic data association filter for cavity boundary extraction from ultrasound images", *IEEE Transactions on Medical Imaging*, vol. 23, no. 6, pp. 772-784, June 2004.

[25] B.Lloyd, "Computation of the Fundamental Matrix", Dept. of Computer Science, University of North Carolina, Open-source software. Available: http://www.cs.unc.edu/~blloyd/comp290-089/fmatrix/

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

[27] R.M.Murray, Z.Li and S.S.Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, 1994.

[28] T.Oskiper, Z.Zhu, S.Samarasekera and R.Kumar, "Visual odometry system using multiple stereo cameras and inertial measurement unit", *in Prof. of IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, U.S.A., June 2007.

[29] M.Kanbara, T.Okuma, H.Takemura and N.Yokoya, "A stereoscopic video see-through augmented reality system based on real-time vision-based registration", in *Proc. of IEEE Virtual Reality*, pp. 255-262, New Brunswick, USA, March 2000.

[30] W.Hoff and T.Vincent, "Analysis of head pose accuracy in augmented reality", *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 4, pp. 319-334, October 2000.

[31] L.Vacchetti and V.Lepetit and P.Fua, "Fusing online and offline information for stable 3D tracking in real-time", *in Prof. of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 241-248, Madison, WI, June 2003.