Robust 3-D Motion Tracking from Stereo Images: A Model-less Method

Ying Kin YU*, Kin Hong WONG, Siu Hang OR and Michael Ming Yuen CHANG

***Abstract*—Traditional vision-based 3-D motion estimation algorithms require given or calculated 3-D models while the motion is being tracked. We propose a high-speed extended-Kalman-filter-based approach that recovers camera position and orientation from stereo image sequences without prior knowledge as well as the procedure for the reconstruction of 3-D structures. Empowered by the use of trifocal tensor, the computation step of 3-D models can be eliminated. The algorithm is thus flexible and can be applied to a wide range of domains. The twist motion model is also adopted to parameterize the 3-D motion. It is minimal since it only has 6 parameters as opposed to 7 parameters in quaternion and 12 parameters in matrix representation. The motion representation is robust because it does not suffer from singularities as Euler angles. Due to the fact that the number of parameters to be estimated is reduced, our algorithm is more efficient, stable and accurate than traditional approaches. The proposed method has been applied to recover the motion from stereo image sequences taken by a robot and a hand-held stereo rig. The results are accurate compared to the ground truths. It is shown in the experiment that our algorithm is not susceptible to outlying point features with the application of a validation gate.**

***Index Terms:* Trifocal tensor, Kalman filtering, Pose tracking, Stereo vision, Structure and motion, Robot Vision, Twist, Visual servoing.**

I. Introduction

One of the most challenging tasks in computer vision is the estimation of position and orientation (pose) from 2-D images. Pose estimation algorithms are useful for a wide range of applications, such as augmented reality, visual servoing and human computer interaction, and can be classified into two major streams. Model-based approaches, in which the exact 3-D structure of the object being tracked must be known, have been widely adopted in the past decade [1][2]. High accuracy can be achieved with the model-based methods but they are restricted to be used under a controlled environment. The second class of approaches take advantages of the structure from motion (SFM) algorithms [3][4]. As a relatively large number of parameters are estimated in a recursive fashion, they are less accurate and stable than the former ones. Usually, prior information [5] and measurements from different types of sensors, such as accelerometer [6], are incorporated to improve the robustness when these methods are applied to robotics. In this paper, a novel pose tracking algorithm is presented. The proposed approach is unique in a way that pose information can be recovered directly from stereo image sequences without the step of reconstructing the 3-D models. It is as accurate, fast and stable as the model-based approaches and its application domain is as wide as the SFM algorithms.

A. Related work

1) Model-based pose estimation

A popular model based pose estimation approach is an iterative method proposed by Lowe [7]. Genetic algorithms have also been used by Hati and Sengupta[8] and further improved by Yu *et. al.* [9]. It incorporates a mismatch filtering strategy into the genetic algorithm using composite chromosomes. Kalman filtering [10] is useful if image sequences, not individual images, are considered. The work [1] uses the extended Kalman filter (EKF) to find the pose of an object based on its CAD model recursively, which can work in real time for servo robot manipulators. Later Lippiello *et. al.* extended it to recover the pose information from stereo image sequences [2].

2) Structure-from-motion-based (SFM) methods

SFM techniques are more general and flexible because it does not require 3D models of the objects. Traditional SFM approaches, such as multiple view geometry [11][12], factorization [13] and bundle adjustment [14][15], tackle the problem by considering the images a batch. This may increase latency, thus less suitable for interactive applications. SFM-based pose tracking algorithms with high-speed and low latency rely on the use of Kalman filters. The series of methods in [16] [17] [3] [5] [6] recover both the structure and motion simultaneously with Kalman filters based on the seminal work by Broida *et. al.* [3]. They applied a single full covariance IEKF to recover the structure and pose of an object. An extension was made by Azarbayejani and Pentland [17] to recover the focal length of the camera in addition to the pose and structure. The pointwise structure is represented by one parameter per point such that the computation is over-determined at every frame when the number of features is larger than 7, resulting in better convergence and stability of the filter. Soatto *et. al.* [18] applied the essential constraint in

This work was supported by a grant (Project No.: 4204/04E) from the Research Grant Council of Hong Kong Special Administrative Region and a direct grant (Project Code: 2050350) from the Faculty of Engineering of the Chinese University of Hong Kong.

Y.K.Yu, K.H.Wong and S.H.Or are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: {ykyu, khwong, shor}@cse.cuhk.edu.hk. Phone:+852-26098438; +852-26098397 ; +852-31634261. Fax:+852-26035024.

Michael Ming Yuen Chang is with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: mchang@ie.cuhk.edu.hk Phone:+852-26098347 Fax:+852-26035032

* The corresponding author

epipolar geometry to Kalman-filter-based motion estimation so that the pose sequence can be computed directly from images. However, the essential matrix becomes degenerate under some commonly appeared motions in real-life [11].

3) Our previous approaches

We developed a 2-step EKF [23] in which the steps of 3-D model recovery and motion estimation are interleaved. And the computation efficiency is increased as a tradeoff in accuracy. We then extended our work in [19] by adding the Interacting Multiple Model into the original formulation. The computation speed is improved and the accuracy is not sacrificed compared to the traditional full covariance EKFs. These two methods are different from the present approach in a way that the simultaneous recovery of motion and 3-D structure is required. Recently, a recursive method based on the trifocal tensor has been proposed [20]. It deals with monocular instead of stereo image sequences, and is in theory less stable than the current version since the use of a tensor in a monocular configuration is not as stable as in a stereo vision setup.

B. Advantages of the proposed algorithm

Recovery of pose without the explicit reconstruction of 3-D models. Neither known 3-D structure nor its computation is required while recovering the pose information from stereo image sequences in the proposed algorithm. Such a characteristic is achieved by the use of trifocal tensor [11], [22] in the Kalman filtering formulation.

High accuracy and computation efficiency. The proposed approach makes use of one EKF that estimates the 3-D motion based on a 6×1 state vector [4]. Also the twist motion model [21] is used to keep track of the pose information, in company with the trifocal tensor and extended Kalman filter, because of the small size of the parameter set, the operation is made to be more efficient.

C. Structure of this article

The organization of this paper is as follows. In Section II, the problem of pose estimation is defined. The geometry of the imaging system and the twist motion model are introduced. In Section III, an overview of the proposed pose tracking algorithm is given. Procedures for the matching of stereo correspondences are outlined. In Section IV, the details of the extended Kalman filter formulation are presented. In Section V, an empirical comparison among our approach, the EKF by Azarbayejani and Pentland [17] and the 2-step EKF by Yu *et. al.* [23] is made using synthetic data. In addition, the proposed method is tested with real image sequences and the results are compared with the ground truths. In Section VI, performance of our approach are discussed.

II. Problem Modeling

A. The imaging system

Fig. 1 shows the geometric setup of the imaging system. The relationships between a point in the 3-D structure and its projection on the left and right image plane are expressed respectively as:

$$\begin{bmatrix}\tilde{u}_{m,t}\\ \tilde{v}_{m,t}\\ \tilde{w}_{m,t}\end{bmatrix} = K[I_{3\times3} \quad 0_{3\times1}]M_t\begin{bmatrix}x_m^W\\ y_m^W\\ z_m^W\\ 1\end{bmatrix}, \qquad \begin{bmatrix}\tilde{u}'_{m,t}\\ \tilde{v}'_{m,t}\\ \tilde{w}'_{m,t}\end{bmatrix} = KEM_t\begin{bmatrix}x_m^W\\ y_m^W\\ z_m^W\\ 1\end{bmatrix} \quad (1)$$

where $X_m^W = [x_m^W \quad y_m^W \quad z_m^W]^T$ denotes the coordinates of the m^{th} model point with respect to the world coordinate frame. K is a 3×3 matrix that encodes the intrinsic parameters of a camera such as the focal length f. E is a 3×4 matrix representing the rigid transformation between the two cameras in the stereo system. K and E are fixed and can be found through the camera calibration process [24]. M_t is a 4×4 matrix having 6 degrees of freedom. It transforms the 3-D structure from the world frame to the reference camera at time instance t. The actual image coordinates $p_{m,t} = [u_{m,t}, v_{m,t}]^T$ on the left view and $p'_{m,t} = [u'_{m,t}, v'_{m,t}]^T$ on the right view are respectively given by

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

B. The twist motion model

Twist is used to parameterize the 3-D pose in our Kalman-filter-based pose tracking algorithm. By definition, a twist can be expressed either as 1) a 6-dimensional vector denoted by ξ_t or 2) a 4×4 matrix denoted by $\tilde{\xi}_t$

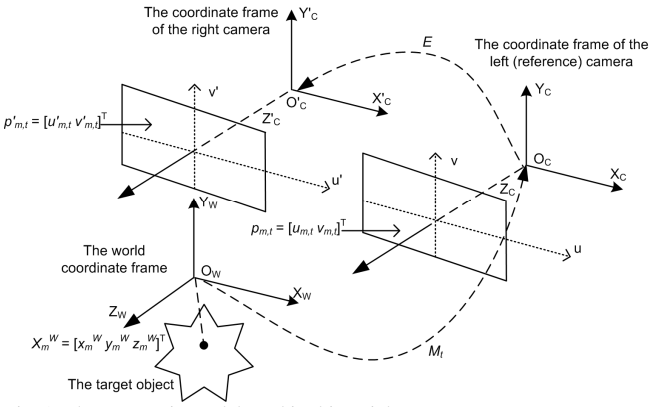

Fig. 1. The geometric model used in this article.

$$\xi_t = \begin{bmatrix}x_t\\ y_t\\ z_t\\ \alpha_t\\ \beta_t\\ \gamma_t\end{bmatrix}, \quad \tilde{\xi}_t = \begin{bmatrix}0 & -\gamma_t & \beta_t & x_t\\ \gamma_t & 0 & -\alpha_t & y_t\\ -\beta_t & \alpha_t & 0 & z_t\\ 0 & 0 & 0 & 0\end{bmatrix} \quad (3)$$

x_t, y_t and z_t are respectively the translations in the x, y and z direction. $\alpha_t, \beta_t, \gamma_t$ are respectively the rotations about the x, y and z axis. More details on the geometric interpretation of twist can be found in [21]. With the exponential map, a twist can be related to the conventional rigid transformation matrix M_t in equation (1)

$$M_t = e^{\tilde{\xi}_t} = I + \tilde{\xi}_t + \frac{(\tilde{\xi}_t)^2}{2!} + \frac{(\tilde{\xi}_t)^3}{3!} + \ldots \quad (4)$$

The twist ξ_t, or equivalently the matrix M_t, encodes the pose information. The objective of the proposed pose tracking algorithm is to compute the object motion, i.e. ξ_t and M_t, at each time-step recursively given only the image measurements $p_{m,t}$ and $p'_{m,t}$.

III. Outline of the Algorithm

A. Feature tracking and extraction

Fig. 2 is an outline of the proposed pose tracking algorithm. The Kanade-Lucas-Tomasi (KLT) tracker [25] is used to extract feature points and track them in the left and right image independently. We assume that the point features extracted by the tracker are contaminated only by Gaussian noise. Their stereo correspondences are setup afterwards to allow a portion of the outliers from the feature tracker to be filtered off, since mis-tracked points in one image are unable to have correspondences in the complementary image of the stereo pair.

B. Setting up stereo correspondences

Features extracted from the stereo image pair (the extrinsic parameters E of the stereo system can be calibrated by a tool [24]) are matched with each other in each time-step. In practice, it is more convenient to find point matches in a stereo pair by estimating the fundamental matrix F directly from the features extracted. In our implementation, F is computed while stereo matching is performed. Features on the left and right images are first matched putatively based on their normalized correlations. The initially matched points are then used to calculate F by the eight-point algorithm [11], together with the Random Sample Consensus (RANSAC) robust estimator [26].

With F, a guided search on the stereo correspondences can be performed. In short, the distance between the m^{th} point in the right view and the epipolar line of the n^{th} point in the left view is $D_{m,n}$ in equation 5.

$$D_{m,n} = \begin{bmatrix} u'_{m,t} \\ v'_{m,t} \\ 1 \end{bmatrix}^T F \begin{bmatrix} u_{n,t} \\ v_{n,t} \\ 1 \end{bmatrix} \tag{5}$$

The pair of points having the smallest $D_{m,n}$ and the highest correlation value is considered as a match. The set of newly acquired matches can be used to improve the accuracy of F until no more matches can be found, see [27] for details. As the intrinsic parameters K of the cameras are known, the required extrinsic parameters E of the stereo system can be found from F [11]. F is required only in the initialization because the relative position of the stereo cameras is fixed here. For all the image pairs at time-step t >1, correspondences are found in real time simply using the guided search discussed in *III-B*

C. Pose tracking

Two trifocal tensors are used to constrain the 2-D positions of the point features in every three views in the measurement model. The first stereo image pair in an image sequence is set as the base pair. It constitutes the first two views of the two trifocal tensors. The third view that builds up the first trifocal tensor T is the image captured by the left camera at time-step t. Similarly, the third view for the second tensor T' is the image taken by the right camera at time-step t. A graphical illustration of the arrangement is shown in Fig. 3.

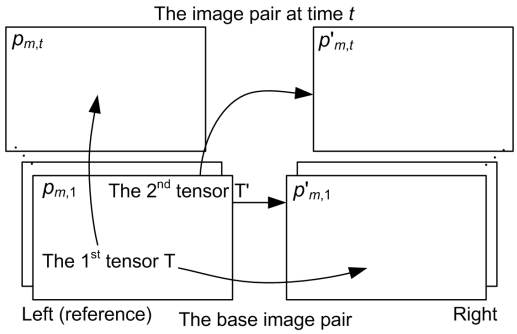

Fig. 3. An illustration of the application of trifocal tensors in the stereo system. The first tensor T involves points $p_{m,1}$, $p'_{m,1}$ and $p_{m,t}$. The second tensor T' involves points $p_{m,1}$, $p'_{m,1}$ and $p'_{m,t}$.

Measures are taken to deal with the changes of the set of observable feature points due to occlusion in real situations. Feature points that can be observed from the set of four views related by the two trifocal tensors are input to the EKF as the measurements. If available point features are below 7, the views at the current time-step will be set as the new base frame pair and the tracker will be bootstrapped. With 7 or more point correspondences across 3 views, the trifocal constraint is able to characterize the rigid motion of the camera. Note that the fundamental matrix F is fixed since the stereo rig is rigid. The treatments in handling occlusions and new point features are relatively simple compared to the existing SFM algorithms.

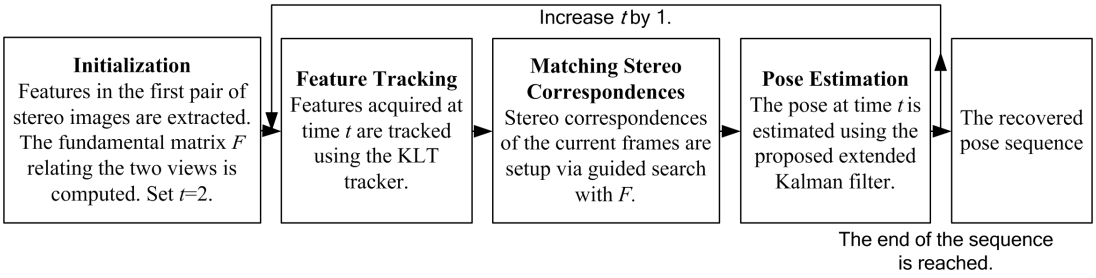

Fig. 2. An outline of the proposed pose tracking algorithm.

IV. Pose Tracking Using EKF and Trifocal Tensors

An extended Kalman filter (EKF) is used to estimate the velocity of the target object in our algorithm. For a stereo sequence having Γ image pairs, the motion of the whole sequence can be recovered after $\Gamma - 1$ Kalman filtering cycles. To make the presentation clear, it is assumed that the point features are observable throughout the sequence so that resetting of the base frame pair is not considered in this section. The formulation of our EKF is as follows. The state vector $\dot{\xi}_t$ is defined as:

$$\dot{\xi}_t = \begin{bmatrix} \dot{x}_t & \dot{y}_t & \dot{z}_t & \dot{\alpha}_t & \dot{\beta}_t & \dot{\gamma}_t \end{bmatrix}^T \tag{6}$$

$\dot{x}_t, \dot{y}_t, \dot{z}_t$ are the translational velocities of object along the x, y and z axis, respectively. $\dot{\alpha}_t, \dot{\beta}_t, \dot{\gamma}_t$ are respectively the angular velocities of object rotation on the x, y and z axis. The dynamic system equations of the filter are as follows:

$$\begin{aligned} M_t &= M_{t-1} \exp(\dot{\xi}_t) \\ \dot{\xi}_t &= \dot{\xi}_{t-1} + \eta_t \end{aligned} \tag{7}$$

The physical meaning of the above equations is that the pose M_t is an integral of velocity while the velocity is an integral of acceleration. The acceleration is modeled as zero-mean Gaussian noise η_t in the EKF. Assuming that the sampling rate of the measurements is high, the motion of the object between successive images in a sequence is small and so do the values of the terms in the velocity vector $\dot{\xi}_t$. The exponential map of $\dot{\xi}_t$ in equation (4) can be approximated by the first order Taylor expansion such that

$$M_t = M_{t-1}(I + \tilde{\dot{\xi}}_t) \tag{8}$$

where $\tilde{\dot{\xi}}_t$ is the matrix form of $\dot{\xi}_t$. The measurement model, which relates the pose M_t and the measurements ε_t acquired from the pair of stereo cameras, is defined as:

$$\varepsilon_t = g_t(M_t) + v_t \tag{9}$$

where v_t is a $4N \times 1$ vector representing zero-mean Gaussian noise imposed on the images captured. Here N is the number of point features extracted from the object being tracked. $g_t(M_t)$ is the $4N \times 1$ -output trifocal tensor point transfer function. Using the image measurements from the first stereo image pair in the sequence, the pose information M_t, together with the extrinsic parameters of the stereo rig E, the estimated coordinates of the feature points at current time t can be computed as:

$$\begin{aligned} g_t(M_t) = \lfloor & u_{1,t} \quad v_{1,t} \quad \ldots \quad u_{m,t} \quad v_{m,t} \quad \ldots \quad u_{N,t} \quad v_{N,t} \\ & u'_{1,t} \quad v'_{1,t} \quad \ldots \quad u'_{m,t} \quad v'_{m,t} \quad \ldots \quad u'_{N,t} \quad v'_{N,t} \rfloor^T \\ [U_{m,t}]^k = [U_{m,1}]^i [l'_m]_j \mathrm{T}_i^{jk}, & \; [U'_{m,t}]^k = [U_{m,1}]^i [l'_m]_j \mathrm{T}'^{jk}_i \end{aligned} \tag{10}$$

The above formulae are written in the tensor notation. $U_{m,1}$, $U_{m,t}$ and $U'_{m,t}$ are respectively the normalized homogenous form of $p_{m,1}$, $p_{m,t}$ and $p'_{m,t}$ such that $U_{m,1} = \begin{bmatrix} \bar{u}_{m,1} & \bar{v}_{m,1} & 1 \end{bmatrix}^T = \begin{bmatrix} \frac{u_{m,1}}{f} & \frac{v_{m,1}}{f} & 1 \end{bmatrix}^T$, $U_{m,t} = \begin{bmatrix} \bar{u}_{m,t} & \bar{v}_{m,t} & 1 \end{bmatrix}^T = \begin{bmatrix} \frac{u_{m,t}}{f} & \frac{v_{m,t}}{f} & 1 \end{bmatrix}^T$ and $U'_{m,t} = \begin{bmatrix} \bar{u}'_{m,t} & \bar{v}'_{m,t} & 1 \end{bmatrix}^T = \begin{bmatrix} \frac{u'_{m,t}}{f} & \frac{v'_{m,t}}{f} & 1 \end{bmatrix}^T$. T is known as the trifocal tensor, which encapsulates the geometric relations among three views [11]. It is analogous to the essential matrix, which is the intrinsic geometry relating two views. By definition, corresponding points in three views have the following relation:

$$[U'_{m,1}]_{\mathrm{X}} \sum_i ([U_{m,1}]^i \mathrm{T}_i)[U_{m,t}]_{\mathrm{X}} = 0_{3x3} \tag{11}$$

With the normalization of the 2-D coordinates, T and T' can be expressed in tensor notation as:

$$\mathrm{T}_i^{jk} = a_i^j a'^k_4 - a_4^j a'^k_i, \qquad \mathrm{T}'^{jk}_i = a''^j_i \, a'^k_4 - a''^j_4 \, a'^k_i \tag{12}$$

a_i^j, a'^j_i and a''^j_i are respectively the elements of the upper 3×4 component of the rigid transformation matrix M_t, the extrinsic parameters E of the stereo system and the matrix product EM_t such that $[I_{3\times3} \quad 0_{3\times1}] M_t = [a_i^j]$, $E = [a'^j_i]$ and $EM_t = [a''^j_i]$. In equation (10), l'_m is a line passing through the m^{th} feature point in the 1st image of the right view, i.e $p'_{m,1}$, and is computed as:

$$\begin{aligned} l'_m &= \begin{bmatrix} \dot{l}_2 & -\dot{l}_1 & -\bar{u}'_{m,1}\dot{l}_2 + \bar{v}'_{m,1}\dot{l}_1 \end{bmatrix}^T \\ \dot{l}'_m &= \begin{bmatrix} \dot{l}_1 & \dot{l}_2 & \dot{l}_3 \end{bmatrix}^T = e_{12} \times U'_{m,1} \end{aligned} \tag{13}$$

where e_{12} is an estimate of the epipole observed from the right camera and is calculated in the initialization step. l'_m is constructed by first finding the epipolar line $\dot{l}'_m$ through the coordinates $U'_{m,1}$. Then l'_m is joining $U'_{m,1}$ and perpendicular to the epipolar line.

The four core Kalman filtering equations can be derived according to the dynamic system (7) and measurement model (9). The prediction equations for calculating the optimal estimates are:

$$\begin{aligned} \hat{\dot{\xi}}_{t,t-1} &= \hat{\dot{\xi}}_{t-1,t-1} \\ P_{t,t-1} &= P_{t-1,t-1} + Q_t \end{aligned} \tag{14}$$

The update equations for the corrections of estimates are:

$$\begin{aligned} \hat{\dot{\xi}}_{t,t} &= \hat{\dot{\xi}}_{t,t-1} + W(\varepsilon_t - g_t(M_t)) \\ P_{t,t} &= P_{t,t-1} - W \nabla g_M P_{t,t-1} \\ W &= P_{t,t-1} \nabla g_M{}^T (\nabla g_M P_{t,t-1} \nabla g_M{}^T + R_t)^{-1} \end{aligned} \tag{15}$$

$\hat{\dot{\xi}}_{t,t-1}$ and $\hat{\dot{\xi}}_{t,t}$ are the estimates of state $\dot{\xi}_t$ after prediction and update, respectively. $P_{t,t-1}$ and $P_{t,t}$ are 6×6 matrices, which are respectively the covariances of $\hat{\dot{\xi}}_{t,t-1}$ and $\hat{\dot{\xi}}_{t,t}$. R_t and Q_t are the covariances of the noise terms v_t and η_t, respectively. W is the $6 \times 4N$ Kalman gain matrix for the filter. ∇g_M is the Jacobian of the non-linear observation equation $g_t(M_t)$ evaluated at $\hat{\dot{\xi}}_{t,t-1}$.

To make the EKF more robust to outliers, a validation gate is added to exclude outlying point features while filtering is performed. Given $S_{m,t}$ is the residual covariance of the measurement pair $p_{m,t}$ and $p'_{m,t}$, the volume of the validation region $V_{m,t}$ of the point pair is defined as:

$$V_{m,t} = G^2 \pi \left| S_{m,t} \right|^{\frac{1}{2}} \quad (16)$$

where $S_{m,t}$ is a 4×4 matrix. G is referred to as the standard deviation of the gate and is a parameter that can be obtained from the chi-square distribution tables in [28]. It is set to 4 in our implementation. Equation (16) is exactly the same as the condition that the measurement pair, so as to be validated, should satisfy the following inequality:

$$r_{m,t}{}^T S_{m,t}^{-1} r_{m,t} < G^2 \quad (17)$$

$$r_{m,t} = \varepsilon_{m,t} - g_{m,t}(M_t) \quad (18)$$

where $\varepsilon_{m,t}$ and $g_{m,t}(M_t)$ are 4×1 vectors representing the 2-D coordinates of a feature pair acquired from the KLT tracker and the coordinates predicted (transferred) by the trifocal tensor, respectively. $r_{m,t}$ is the innovation of an individual measurement pair $p_{m,t}$ and $p'_{m,t}$. Only the validated point features are used for the corrections of the state estimates.

V. Experiments and Results

A. *Experiments with synthetic data*

Synthetic structure points, centered at 0.5m away from the camera, were generated. The motion parameters were randomly set from 0.2 to 1.2 degrees per frame for each rotation angle and 0.005 to 0.015 meters per frame for each translation parameter. The focal length of the camera was 6mm with a 2-D zero-mean Gaussian noise of 1 pixel standard deviation. The length of each synthetic sequence was 99 frames. In the simulation, the two cameras in the stereo system were placed 0.05m apart. They were pointing towards the positive direction of the z-axis. A sample of the synthetic images showing the feature distribution can be found in Fig. 4.

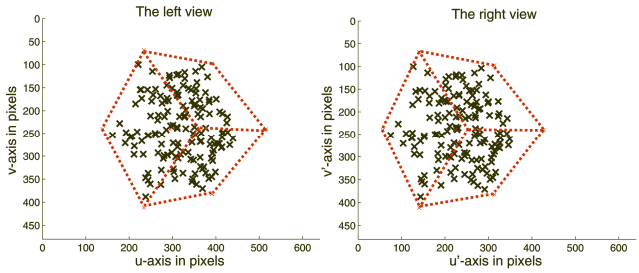

Fig. 4. A sample of the synthetic stereo images at time t=1. The left and right plot show the optical distribution of the point features (indicated by cross markers) in left and right view, respectively. The dotted lines illustrate the cube that virtually present in the space bounding the point features.

The proposed algorithm, the EKF by Azarbayejani and Pentland [17] and the 2-step EKF by Yu *et. al.* [23] are compared. Fig. 5 shows the percentage of accumulated total errors, the lines with asterisk (*), triangle (△) and circle (○) markers represent the proposed approach, the EKF by Azarbayejani and Pentland [17] and the 2-step EKF by Yu *et. al.* [23], respectively. To make the average values more meaningful, the values for the diverged cases were excluded in plotting these graphs. The average percentages of accumulated total rotation and translation errors are respectively defined as the average of the differences between the actual and the recovered angles in the axis-angle representation, and the average of the absolute differences between the actual and the recovered translations in terms of percentages. From Fig. 5, the proposed approach had the lowest errors for the whole range of number of point features in the experiment. By averaging a total of 750 independent test cases of 15 different conditions, it achieved an average total rotation and translation errors of 0.6010 degrees and 0.0149 meters, respectively.

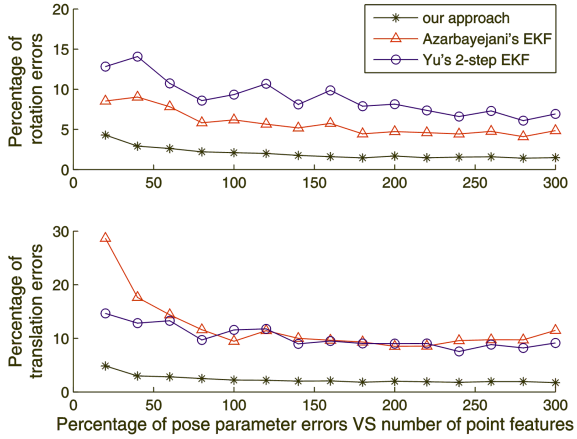

Fig. 5. The average percentages of accumulated rotation errors (top) and translation errors (bottom) versus the number of point features input to the 3 algorithms. The diverged cases were removed before computing the values.

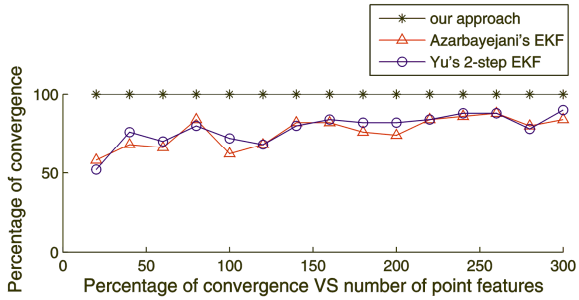

Fig. 6. The percentage of convergence of the 3 algorithms against the number of point features.

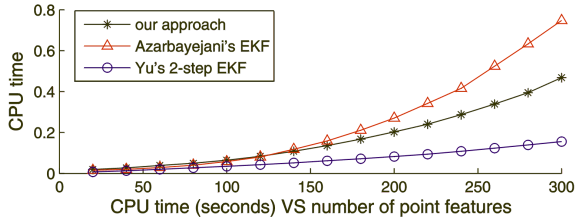

Fig. 7. The CPU (in Matlab and runs on a Pentium IV 2GHz) time required by the 3 algorithms at each time-step against the number of point features

Fig. 6 illustrates the relationship between the percentage of convergence and the number of point features, which is actually an indicator of the algorithm stability. In general, the stability was improved as a larger number of corner features were input to the filters. Our novel approach achieved a hundred percent of convergence. The processing time of the algorithms versus the number of measurements is shown in Fig. 7.

B. *Experiments with real images*

A real stereo image sequence captured by cameras on a computer controlled robot was tested. The sequence has 115 frames and the robot was driven in front of a bookshelf in an office. The robot was driven by two stepping motors, which were linked to the wheels on each side of the robot. The motors received control codes from a personal computer. Steering could be achieved by applying unequal inputs to the left and right motor. For example, an input of one unit to the left motor and a negative step to the right motor would result in turning right to a certain degree. Two digital cameras were mounted on the top of the robot and pictures were transmitted back via the Universal Serial Bus (USB). To compare the recovered pose with the ground truth, both the robot and cameras were required to be calibrated. The diameters of the wheels and distance between them were first measured. Then the robot displacement per motor step was found out. Together with the measurements of the wheel arrangement, the rotation of the robot for one step difference between the left and right motor could be calculated.

TABLE I
THE INTRINSIC PARAMETERS OF THE CAMERAS

	The left camera	The right camera
Focal length (f_x, f_y) in pixels	(409.09, 408.40)	(403.28, 402.80)
Coordinates of the principal point (s_x, s_y)	(162.67, 118.87)	(162.07, 118.28)

A table showing the intrinsic parameters of the cameras in the experiment.

The intrinsic parameters of the cameras were calibrated by the tool in [24]. 30 images of a planar checkerboard pattern with known dimensions were used. The calibration details, such as the focal length and principal point position, of the two cameras are listed in Table I. The readers can refer to [24] for the procedures involved. Actually, deviations of camera parameters from the actual values mainly affect the precision of the point transfer function in the measurement model (equation 10) that corrects the prediction of state estimates. Over-estimation of the focal length makes the recovered rotation angles smaller than the real ones due to the reduction in field of view.

In Figs. 8 and 9, we want to show whether the 3-D motion of the robot can be correctly found by our method. First, corner features in the 1st image pair of the sequence were extracted. Then, a set of trifocal tensors was computed using the motion parameters recovered and was used to transfer (re-project) the corner features from the 1st to the succeeding image pairs. The consistency of the motion of these corner features with respect to the background images was checked.

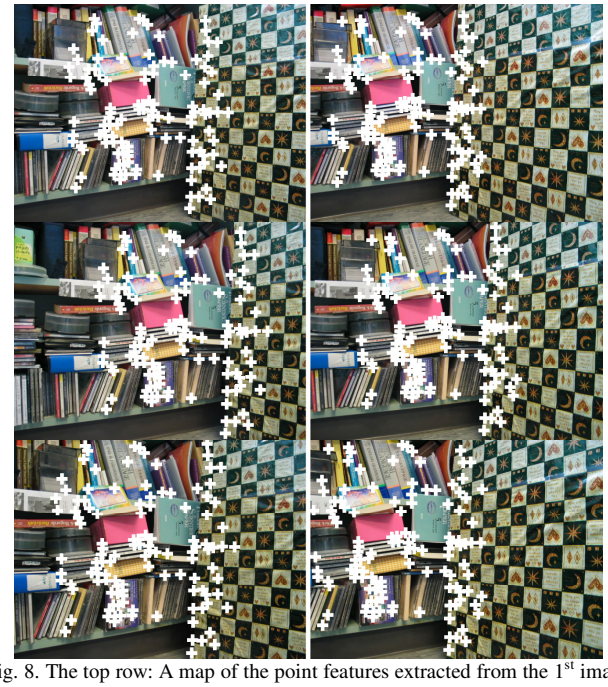

Fig. 8. The top row: A map of the point features extracted from the 1st image pair of the stereo image sequence. The middle row: The re-projections of the extracted point features in the 35th image pair. The bottom row: The re-projections of the point features in the 100th image pair. Note that the left and the right column show the images captured by the left and right camera in the stereo system, respectively.

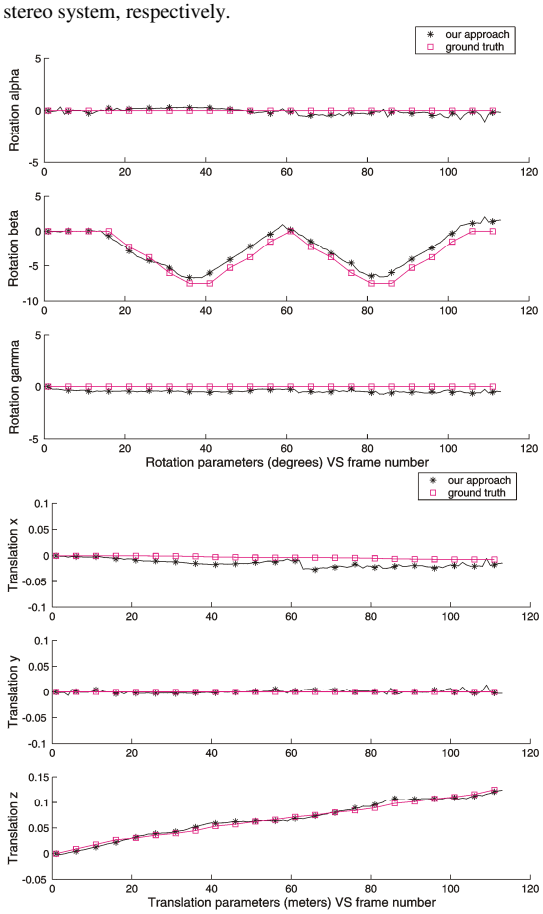

Fig. 9. A comparison of the pose recovered from the real stereo image sequence with the ground truth.

Fig. 8 shows the results of re-projecting the features from the 1st to the 35th and 100th stereo pair with the recovered 3-D motion. It can be observed that the features, which are indicated by plus signs (+), stick to the same position relative to the background in all the three pairs of images. A video demonstrating the results can be found at the URL http://www.cse.cuhk.edu.hk/~vision/. In addition, the recovered pose sequence was compared with the ground truth, as indicated by the lines with square markers (□) in Fig. 9. The average total rotation and translation errors were respectively 0.8773 degrees and 0.0125 meters.

VI. Comparisons with Existing Algorithms Tested in the Experiments

A. Algorithm design, accuracy and stability

The trifocal tensor expresses the geometric relations among point correspondences in three views. By locking every three views in the stereo system with the trifocal tensor point transfer function, it is adequate for our EKF to compute the pose *directly* from images and the explicit reconstruction of 3-D structure is not necessary. The essential matrix in epipolar geometry, which is analogous to the trifocal tensor, relates two instead of three views. Applying the epipolar constraint, which arises from the essential matrix, to the EKF is possible and also has the advantage of eliminating the 3-D structure recovery step. However, the essential matrix is susceptible to degeneracy in real situations, such as observing critical surfaces or undergoing a pure rotation camera motion [11]. With the assumption that the scene is a non-planar rigid body, the use of trifocal constraint in a *stereo configuration* is more robust and has almost no restrictions on the camera motion to be estimated.

SFM-based pose tracking algorithms, for example the approaches [16] [23] tested in the previous section, should have a lower accuracy and stability than the proposed EKF. The reason is that these traditional algorithms require the recursive optimization of a large number of parameters. Our new approach, which has no dependency on the 3-D structure, only needs to compute the six velocity parameters in a Kalman filtering cycle.

The experiments show that our algorithm reduced the errors by at least 60% and was the most stable under the experimental conditions. The proposed approach reached a hundred percent of convergence even if the image noise was as high as 1 pixel standard deviation. On the other hand, the EKF by Azarbayejani and Pentland and the 2-step EKF by Yu *et. al.* only had the convergence probabilities of 76.13% and 78.27%, respectively. Our method had a higher convergence rate than the two other algorithms mainly because of the application of the trifocal constraint to the motion recovery process, making the approximation of the non-linear measurement model in the EKF more robust.

B. Computation speed

In the proposed approach, there is one 6×1 state vector for the estimation of pose in the EKF. Its size is independent from N, where N is the number of point features input to the filter. On the other hand, traditional recursive SFM approaches that are capable of recovering 3-D motion from an unknown environment, for example the EKF by Azarbayejani and Pentland [16] tested in the experiment, compute the pose from an image sequence with a more complex EKF based on a $(N+7)\times1$ state vector. The dependence of the size of the state vector on N results in a larger number of computation steps even for a moderate value of N compared to our method.

The 2-step EKF by Yu *et. al.* [23], which is a decoupled EKFs for the SFM problem, consists of $N + 1$ small filters, in which N of them employ a 3×1 state vector to compute the structure and the rest one uses a 12×1 state vector to estimate the pose. It has a higher speed than our method mainly because there is no additional computation overhead in constructing the trifocal tensor point transfer function. However, it has a lower accuracy due to the decoupling of filters.

From the experimental results, our algorithm took 0.175s to compute the pose from an additional image pair on average. It outperformed the full covariance EKF by Azarbayejani and Pentland, which needed 0.244s to recover the pose from a single image having the same number of measurements. As Yu's 2-step EKF is decoupled to make a tradeoff between speed and accuracy, it is reasonable that it took 0.069s to process an extra image frame.

C. Handling occluded point features

As point features are input to our EKF as measurements, adding or removing them from the EKF can be arbitrary, provided that the number of features passed to the filter is greater than or equal to 7. The procedures are much more complicated for the SFM-based pose tracking methods. Due to the fact that the final pose sequence is highly dependent on the correctness of the recovered structure, keeping track of the corresponding features in the 3-D structure properly is crucial in traditional SFM algorithms, thus increasing the difficulties in implementation.

VII. Conclusion

A novel recursive pose tracking algorithm based on stereo vision is proposed in this article. The trifocal tensor has been incorporated into the measurement model of the extended Kalman filter to constraint every three views in the image sequence. Pose estimation no longer depends on the 3-D structure and updating of the structure is not necessary while estimating the pose. Outlying point features caused by the feature tracker are removed in the guided search when the stereo correspondences are setup. The validation gate embedded in the EKF filters off any remaining outliers due to point mismatches in each pair of stereo image. The twist motion model is adopted to make the motion representation robust and minimal. The proposed approach has a high performance in speed, accuracy, stability, together with a simple procedure to handle occluded point features. Even so the application domain is not sacrificed and the full covariance nature of an EKF-based SFM algorithm, in which the pose and the pointwise structure are fully coupled, is kept implicitly. Experimental results show that our approach

outperformed other EKF-based SFM methods. The results of pose tracking using real images were precise compared to the ground truths. We believe the design of the proposed algorithm can definitely enhance the performance of interactive applications like augmented reality, visual servoing and human computer interaction.

REFERENCES

[1] W.J.Wilson, C.C.Williams Hulls and G.S.Bell, "Relative end-effector control using Cartesian position based visual servoing", *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 684-696, October 1996.

[2] 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, September 2001.

[3] 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 1990.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] S.Hati and S.Sengupta, "Robust camera parameter estimation using genetic algorithm", *Pattern Recognition Letters*, vol. 22, no. 3/4, pp. 289-298, March 2001.

[9] Y.K.Yu, K.H.Wong and M.M.Y.Chang, "Pose estimation for augmented reality applications using genetic algorithm", *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 35, no. 6, pp. 1295-1301, December 2005.

[10] M.S.Grewal and A.P.Andrews, *Kalman Filtering Theory and Practice*, Prentice Hall, 1993.

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

[12] 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.

[13] C.J.Poelman and T.Kanade, "A paraperspective factorization method for shape and motion recovery", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 206-218, March 1997.

[14] B.Triggs, P.McLauchlan, R.Hartley and A.Fitzgibbon, "Bundle adjustment – A modern synthesis", in *Proc. of the International Workshop on Vision Algorithms: Theory and Practice*, pp. 298-372, Corfu Greece, September 1999.

[15] Z.Zhang and Y.Shan, "Incremental motion estimation through modified bundle adjustment", in *Proc. of the IEEE International Conference on Image Processing*, vol. 2, pp.343-346, Barcelona, September 2003.

[16] 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.

[17] 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.

[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] 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.

[20] 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 Systems, Man and Cybernetics, Part B: Cybernetics*, vol. 36, no. 5, pp. 1081- 1090, October 2006.

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

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

[23] 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.

[24] J-Y.Bouguet, "Camera Calibration Toolbox for Matlab", Dept. of Electrical Engineering, California Institute of Technology, Open-source software. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/

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

[26] 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.

[27] 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/

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

Ying Kin Yu received a B.Eng (First Class Honours), an M.Phil. and a Ph.D. degree from the Chinese University of Hong Kong in 2002, 2004 and 2007, respectively. He has been awarded the Sir Edward Youde Memorial Fellowship twice for his academic achievements. His research interests are computer vision, image processing, Kalman filtering and genetic algorithms.

Kin Hong Wong received a Ph.D. from the University of Cambridge, U.K in 1986. He is now an Associate Professor at the Computer Science and Engineering Department of the Chinese University of Hong Kong

Siu Hang Or received a Ph.D. from the Chinese University of Hong Kong in 1998. He is now an instructor at the Computer Science and Engineering Department of the Chinese University of Hong Kong

Michael Ming Yuen Chang received a Ph.D. degree from University of Cambridge, UK in 1988. He is now an Associate Professor at the Information Engineering Department of the Chinese University of Hong Kong.