In Proc. of The IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06), New York, Jun. 2006.

Recursive Recovery of Position and Orientation from Stereo Image Sequences without Three-Dimensional Structures

Ying Kin Yu, Kin Hong Wong, Siu Hang Or
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong
{ykyu, khwong, shor}@cse.cuhk.edu.hk

Michael Ming Yuen Chang
Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong
mchang@ie.cuhk.edu.hk

Abstract

Traditional vision-based 3-D motion estimation algorithms for robots 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 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 the trifocal tensor, the computation step of 3-D models can be eliminated. The algorithm is thus more flexible and can be applied to a wide range of domains. The twist motion model is also adopted to parameterize the 3-D motion such that the motion representation in the proposed algorithm is robust and minimal. As the number of parameters to be estimated is reduced, our algorithm is more efficient, stable and accurate compared to traditional approaches. The proposed method has been verified using a real image sequence with ground truth.

1. Introduction

One of the most important elements for robot vision is the estimation of position and orientation (pose). Traditional pose estimation algorithms that are useful for robotic applications such as visual servoing and localization are 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 [19] [20] [26]. High accuracy can be achieved with the model-based methods but they are confined to be used under a controlled environment. The second class of approaches take the advantages of the structure from motion (SFM) algorithms in the computer vision community [3] [10] [17] [25]. As a relatively large number of parameters are estimated in a recursive fashion, this class of algorithms is less accurate and stable than the former ones. Usually, prior information [4] or measurements from different types of sensors, such as accelerometer [9], are incorporated to improve the robustness. In this paper, a novel pose tracking algorithm designed for robot vision 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.

1.1. Related work

SFM approaches, such as multiple view geometry [13], factorization [15] and bundle adjustment [14][16], compute structure and motion in a batch. As a number of images are required to be considered at one time, these methods suffer from a certain degree of latency and are less suitable for interactive applications like visual servoing in robotics. SFM-based pose tracking algorithms with high-speed and low latency [1] [2] [3] [4] [5] [6] [7] [8] [9] rely on the use of Kalman filters [11]. Broida *et. al.* [3] applied a single full covariance iterated extended Kalman filter to recover the structure and pose of an object. Azarbayejani and Pentland [2] extended the previous work [3] to recover the focal length of the camera in addition to the pose and structure using an extended Kalman filter (EKF). Also, the 3-D structure is represented by one parameter per point. Yu *et. al.* [6] decoupled the full covariance EKF such that the computation of pose and structure is interleaved. They then extended their work by adding the Interacting Multiple Model into the original formulation [7]. Soatto *et. al.* [18] applied the essential constraint in 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 [13]. Therefore, Yu *et. al.* [8] employed the trifocal constraint to tackle the problem. Similar techniques in SFM have also been applied to simultaneous localization and map-building for robot navigation. The system in [4] uses an active stereo head to acquire image features and the recovered motion is constrained to translation on the x-z plane and rotation on the pitch angle. Constraints on the motion are relaxed in [9] so that localization on undulating terrain is possible. The navigation of the robot is assisted with the use of roll/pitch sensor via an accelerometer sensory mechanism.

Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)
0-7695-2597-0/06 $20.00 © 2006 IEEE

1.2. Advantages of the proposed algorithm

The objective of this article is to present a high-speed, accurate and stable SFM-based pose tracking algorithm without the assistance of odometer or accelerometer. The major advantages of our approach are summarized as follows:

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 the trifocal tensor [13] in the Kalman filtering formulation. Without handling the 3-D models in the filter, the values that are required to be computed in each filtering cycle are limited to the six velocity parameters of the pose only. The number of parameters is as small as the model-based pose estimation method.

Application of the twist motion model in the trifocal tensor. The twist motion model [22] is used to keep track of the pose information, in company with the trifocal tensor and EKF. Compared to the other representation of the pose, such as the direct use of the matrix (12 parameters) or translation plus quaternion (7 parameters), the twist motion model having a total of 6 parameters is minimal. Euler angles and translation vector, which also have 6 elements, can be used to represent the 3-D pose. However, such a parameterization suffers from singularities. The use of the twist motion model to encode the 3-D motion in our algorithm is robust and minimal.

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. The traditional recursive approaches for the SFM problem, for example the one in [2], compute the pose from an image sequence with a more complex EKF based on a $(N+7)\times1$ state vector, where N is the number of point features input to the filter. The computation speed of our method is higher than that of the traditional EKFs since the sizes of the matrices involved in the filtering process are smaller.

As the proposed algorithm makes use of the trifocal constraint [13], which is a constraint in the imaging system, in addition to the dynamic system constraint in the EKF, its estimation accuracy has been improved. It is shown in the experiment that our approach has a better overall performance than other existing methods [2] [6]. The proposed approach has been tested using a real image sequence and the result is accurate compared to the ground truth.

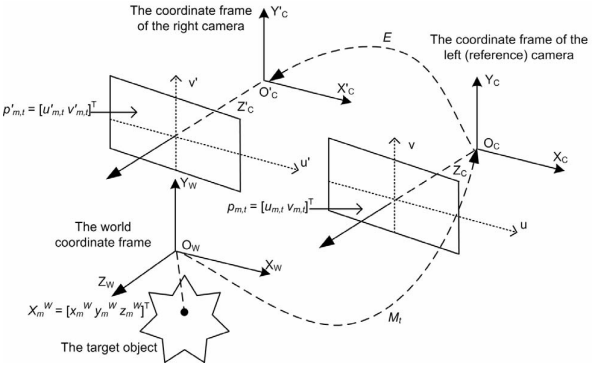

Fig. 1. The geometric model used in this article.

2. Problem Modeling

2.1. The imaging system

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

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

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 and is in the form of

$$K = \begin{bmatrix}-f/s_x & 0 & o_x\\ 0 & -f/s_y & o_y\\ 0 & 0 & 1\end{bmatrix} \quad (2)$$

where f is the focal length. $[o_x \quad o_y]^T$ and $[s_x \quad s_y]^T$ are the coordinates of the image center and the effective size of a pixel, respectively. For simplicity, the two cameras used in our stereo system are assumed to be identical. E is a 3×4 matrix representing the rigid transformation between the two cameras. K and E are fixed and can be found in the camera calibration process [23]. M_t is a 4×4 matrix that transforms the 3-D structure from the world frame to the reference camera at time instance t and can be written as

$$M_t = \begin{bmatrix}R_t & T_t\\ 0\ 0\ 0 & 1\end{bmatrix} \quad (3)$$

where R_t is a 3×3 rotation matrix and T_t is a 3×1 translation vector. M_t has 6 degrees of freedom. 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

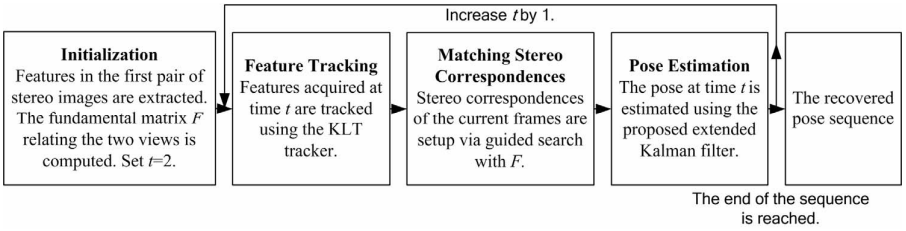

Fig. 2. An outline of the proposed pose tracking algorithm.

$$\begin{bmatrix} u_{m,t} \\ v_{m,t} \end{bmatrix} = \begin{bmatrix} \widetilde{u}_{m,t} / \widetilde{w}_{m,t} \\ \widetilde{v}_{m,t} / \widetilde{w}_{m,t} \end{bmatrix}, \quad \begin{bmatrix} u'_{m,t} \\ v'_{m,t} \end{bmatrix} = \begin{bmatrix} \widetilde{u}'_{m,t} / \widetilde{w}'_{m,t} \\ \widetilde{v}'_{m,t} / \widetilde{w}'_{m,t} \end{bmatrix} \tag{4}$$

2.2. 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 $\widetilde{\xi}_t$

$$\xi_t = \begin{bmatrix} x_t \\ y_t \\ z_t \\ \alpha_t \\ \beta_t \\ \gamma_t \end{bmatrix}, \quad \widetilde{\xi}_t = \begin{bmatrix} 0 & -\gamma_t & \beta_t & x_t \\ \gamma_t & 0 & -\alpha_t & y_t \\ -\beta_t & \alpha_t & 0 & z_t \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{5}$$

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 [22]. With the exponential map, a twist can be related to the conventional rigid transformation matrix M_t in (1)

$$M_t = e^{\widetilde{\xi}_t} = I + \widetilde{\xi}_t + \frac{(\widetilde{\xi}_t)^2}{2!} + \frac{(\widetilde{\xi}_t)^3}{3!} + \ldots \tag{6}$$

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 $p'_{m,t}$

3. An outline of the algorithm

3.1. Feature tracking and extraction

Fig. 2 is an outline of the proposed pose tracking algorithm. The Kanade-Lucas-Tomasi (KLT) tracker described in [12] is used to extract feature points and track them in the images. We assume that the point features extracted by the tracker are contaminated only by Gaussian noise. The features from the left and the right image sequences are tracked independently. Their stereo correspondences are setup afterwards. Such a scheme allows outliers to be filtered off, since mis-tracked points in one image are unable to have correspondences in the complementary image of the stereo pair.

3.2. Setting up stereo correspondences

Features extracted from the stereo image pair are matched with each other in each time-step. To setup the stereo correspondences, we require that the configuration of the stereo system, i.e. the extrinsic parameters E, is known and this can be achieved by calibrating the pair of cameras using the tools in [23]. In practice, the relative position and orientation of the two cameras may be adjusted according to the actual situation. It is not so convenient to compute E using a calibration pattern every time after adjustment.

Another way to find point matches in a stereo pair is 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 [13], together with the Random Sample Concensus (RANSAC) robust estimator [21].

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

$$D_{m,n} = \begin{bmatrix} u'_{m,t} \\ v'_{m,t} \\ 1 \end{bmatrix}^T F \begin{bmatrix} u_{n,t} \\ v_{n,t} \\ 1 \end{bmatrix} \tag{7}$$

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. This procedure is known as iterative improvement and readers

can refer to [24] for details. As the intrinsic parameters K of the cameras are known, the required extrinsic parameters E of the stereo system can be computed from F according to [13].

Since we assume that the cameras are fixed while operating, the computation of F is required only in the initialization. For all the image pairs at time-step t >1, correspondences are found simply using the guided search mentioned before. The refinement of F is not necessary. The time taken for such guided search is fast and feasible for real-time implementation.

3.3. Pose tracking

The extended Kalman filter (EKF), in company with the trifocal tensor, are used to estimate the pose of an object in the image sequences. The EKF models the dynamics of an object as acceleration having a zero-mean Gaussian value. The trifocal tensor constrains the 2-D positions of the point features in every three views in the measurement model. Two tensors are required and they are applied as follows. The first stereo image pair in an image sequence is set as the base pairs. They constitute 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.

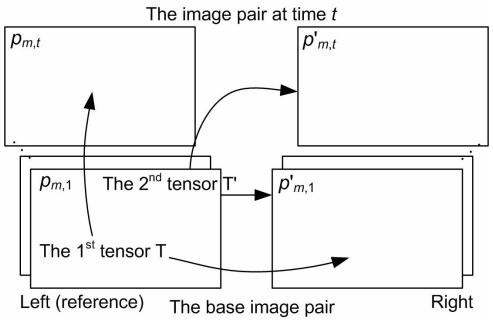

Fig. 3. An illustration of the application of trifocal tensors in the stereo system. The first tensor T involves points $p_{m,1}$, $p'_{m,1}$ and $p_{m,t}$. The second tensor T' involves points $p_{m,1}$, $p'_{m,1}$ and $p'_{m,t}$.

In real situations, features may disappear due to occlusion and new features may appear when the environmental conditions change. Measures are taken to deal with the changes of the set of observable feature points. 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 the number of available point features is 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 for matching stereo correspondences is not required to be re-computed as the relative pose of the stereo pair is unchanged throughout the sequence. The treatments in handling occlusions and new point features are relatively simple compared to the existing SFM algorithms.

4. Pose tracking using the extended Kalman filter and trifocal tensors

At each Kalman filtering cycle of our algorithm, the EKF estimates the velocity of the target object at the current time-step. For the clarity of the presentation, it is assumed that the point features are observable in the whole stereo image 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:

$$\dot{\xi}_t = \begin{bmatrix} \dot{x}_t & \dot{y}_t & \dot{z}_t & \dot{\alpha}_t & \dot{\beta}_t & \dot{\gamma}_t \end{bmatrix}^T \tag{8}$$

$\dot{x}_t, \dot{y}_t, \dot{z}_t$ 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:

$$\begin{aligned} M_t &= M_{t-1} \exp(\dot{\xi}_t) \\ \dot{\xi}_t &= \dot{\xi}_{t-1} + \eta_t \end{aligned} \tag{9}$$

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 the 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 (6) can be approximated by the first order Taylor expansion such that

$$M_t = M_{t-1}(I + \tilde{\dot{\xi}}_t) \tag{10}$$

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:

$$\varepsilon_t = g_t(M_t) + v_t \tag{11}$$

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:

TABLE I
A SUMMARY OF THE ALGORITHM PERFORMANCE

	Our approach	Azarbayejani's EKF	Yu's 2-step EKF
The time required to process an extra image	0.130s	0.150s	0.061s
The average percentage of the accumulated total rotation error (Diverged cases excluded)	1.7837%	5.1451%	9.5453%
The average percentage of the accumulated total translation error (Diverged cases excluded)	2.1181%	11.866%	12.293%
The percentage of convergence	100%	88%	86%

A table showing a summary of the performance of the 3 algorithms under comparison.

$$g_t(M_t) = \begin{bmatrix} u_{1,t} & v_{1,t} & \dots & u_{m,t} & v_{m,t} & \dots & u_{N,t} & v_{N,t} \\ u'_{1,t} & v'_{1,t} & \dots & u'_{m,t} & v'_{m,t} & \dots & u'_{N,t} & v'_{N,t} \end{bmatrix}^T$$

$$[U_{m,t}]^k = [U_{m,1}]^i [l'_m]_j \mathrm{T}_i^{jk},\ [U'_{m,t}]^k = [U_{m,1}]^i [l'_m]_j \mathrm{T}'^{jk}_i \quad (12)$$

The above equations are written in tensor notation. $U_{m,t}$ and $U'_{m,t}$ are respectively the normalized homogenous form of $p_{m,t}$ and $p'_{m,t}$ such that $U_{m,t} = [\bar{u}_{m,t} \quad \bar{v}_{m,t} \quad \bar{w}_{m,t}]^T = [u_{m,t}/f \quad v_{m,t}/f \quad 1]^T$ and $U'_{m,t} = [\bar{u}'_{m,t} \quad \bar{v}'_{m,t} \quad \bar{w}'_{m,t}]^T = [u'_{m,t}/f \quad v'_{m,t}/f \quad 1]^T$. T and T' are known as the trifocal tensor, which encapsulates the geometric relations among three views [13]. The trifocal tensor is analogous to the essential matrix, which is the intrinsic geometry relating two views. By definition, corresponding points in three views, for example $p_{m,1}$, $p'_{m,1}$ and $p_{m,t}$, have the following relation:

$$[U'_{m,1}]_\times \sum_i ([U_{m,1}]^i \mathrm{T}_i)[U_{m,t}]_\times = 0_{3\times 3} \quad (13)$$

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

$$\mathrm{T}_i^{jk} = a_i^j a'^k_4 - a_4^j a'^k_i,\ \mathrm{T}'^{jk}_i = a''^j_i\, a'^k_4 - a''^j_4\, a'^k_i \quad (14)$$

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 $[I_{3\times 3} \quad 0_{3\times 1}]M_t = [a_i^j]$, $E = [a'^j_i]$ and $EM_t = [a''^j_i]$. In (12), 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

$$l'_m = \begin{bmatrix} \dot{l}_2 & -\dot{l}_1 & -\bar{u}'_{m,1}\dot{l}_2 + \bar{v}'_{m,1}\dot{l}_1 \end{bmatrix}^T$$
$$\dot{l}'_m = \begin{bmatrix} \dot{l}_1 & \dot{l}_2 & \dot{l}_3 \end{bmatrix}^T = e_{12} \times U'_{m,1} \quad (15)$$

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 first by 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.

According to the dynamic system (9) and measurement model (11), the core Kalman filtering equations can be derived. The prediction equations for calculating the optimal estimates are

$$\hat{\dot{\xi}}_{t,t-1} = \hat{\dot{\xi}}_{t-1,t-1}$$
$$P_{t,t-1} = P_{t-1,t-1} + Q_t \quad (16)$$

The update equations for the corrections of estimates are

$$\hat{\dot{\xi}}_{t,t} = \hat{\dot{\xi}}_{t,t-1} + W(\varepsilon_t - g_t(M_t))$$
$$P_{t,t} = P_{t,t-1} - W\nabla g_M P_{t,t-1}$$
$$W = P_{t,t-1}\nabla g_M{}^T (\nabla g_M P_{t,t-1} \nabla g_M{}^T + C_t)^{-1} \quad (17)$$

$\hat{\dot{\xi}}_{t,t-1}$ 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 $\hat{\dot{\xi}}_{t,t}$. C_t and Q_t are the covariances of the noise terms ν_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 $\hat{\dot{\xi}}_{t,t-1}$.

5. Experiments and results

5.1. The synthetic data experiment

A synthetic structure with 150 random feature points was generated. The motion of the object was composed of three different segments, a pure translation section, a pure rotation section and a mixed motion section. The motion parameters were generated randomly 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 camera had 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.

The proposed algorithms, the EKF by Azarbayejani and

Pentland [2] and the 2-step EKF by Yu *et. al.* [6] were implemented in Matlab and run on a Pentium IV 2GHz machine to estimate the camera motion. A total of 50 independent tests were carried out. To make a fair comparison, the number of measurements input to the EKFs was equal. It means that our approach only made use of 75, instead of 150, point features, resulting in 150 measurements from a pair of stereo images.

Table I summarizes the overall performance of the three algorithms. The first row shows the computation time needed to recover the camera motion when images were sequentially input to the EKFs. Our algorithm had a higher speed than the full covariance EKF by Azarbayejani and Pentland but lower than that of the 2-step EKF by Yu *et. al.*. The reason is that the 2-step EKF is decoupled, which is actually a tradeoff between speed and accuracy. The second and the third rows are the average percentages of the accumulated total rotation and translation error. The total rotation error is defined as the difference between the actual and the recovered angle in the axis-angle representation. The total translation error is the magnitude of the difference between the recovered translation and the actual one. The average accumulated errors were computed by first summing up the corresponding pose parameter errors of all the image frames. They were then divided by the number of frames in the sequence and the averages of all the test cases were taken. The errors were finally expressed in terms of percentages. Our algorithm reduced the errors by at least 60% and 80% for rotation and translation, respectively. The fourth row lists the percentages of convergence of the algorithms, which can be regarded as an indicator of the algorithm stability. Our method was the most stable under the experimental conditions.

5.2. The real image experiment

An experiment using real scene images was also performed. A stereo image sequence with ground truth was used to test the proposed approach. The sequence was taken by a robot moving along a zigzag path on the floor. The length of the image sequence was 115 frames. Some samples of the images in the sequence are shown in Fig. 4. Fig. 5 are the results of pose tracking. In the tracking process, the number of point features that was able to setup stereo correspondences in an image pair ranged from 15 to 70. The recovered pose was compared with the ground truth, which is indicated by the line with square markers. It was accurate compared to the real values. More results can be found at http://www.cse.cuhk.edu.hk/~khwong/demo/

Fig. 4. Some samples of the pictures in the stereo image sequence. The 1st row: The 35th image pair. The 2nd row: The last (115th) image pair.

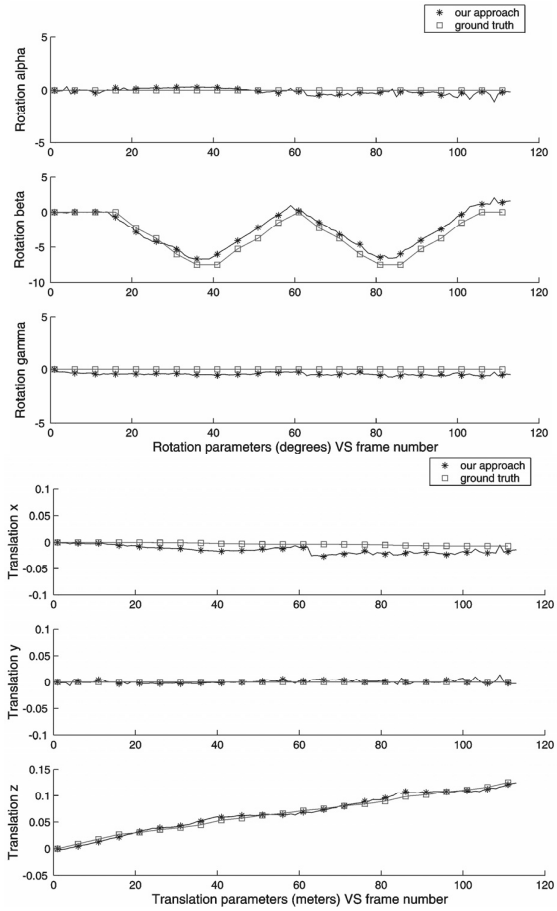

Fig. 5. A comparison of the pose recovered from the real stereo image sequence with the ground truth.

6. Conclusion

A novel EKF-based pose tracking algorithm based on stereo vision is proposed in this article. With the trifocal tensor, 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 in the images are removed in the guided search when the stereo correspondences are setup. 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 the changeable set of point features. Even so the application domain is not sacrificed and the full covariance nature of an EKF-based SFM algorithm, in which the recovery of the pose and the 3-D structure are fully coupled, is kept implicitly. Experimental results show that our approach had a better overall performance than the other EKF-based SFM methods. The results of pose tracking using real images were accurate compared to the ground truth. We believe that the design of the proposed algorithm can satisfy the requirements of robotic applications like visual servoing and localization.

7. Acknowledgement

The work described in this paper 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.

References

[1] A.Chiuso, P.Favaro, H.Jin and S.Soatto, "Structure from motion causally integrated over time", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, no. 4, pp. 523-535, Apr. 2002.

[2] A.Azarbayejani and A.P.Pentland, "Recursive estimation of motion, structure, and focal length", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. 6, pp. 562-575, Jun. 1995.

[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, Jul. 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, Jul. 2002.

[5] J.Weng, N.Ahuja and T.S.Huang, "Optimal motion and structure estimation", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 9, pp. 864-884, Sep. 1993.

[6] Y.K.Yu, K.H.Wong and M.M.Y.Chang, "Recursive three-dimensional model reconstruction based on Kalman filtering", *IEEE Trans. Syst., Man, Cybern. B,* vol. 35, no. 3, pp. 587-592, Jun. 2005.

[7] 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 Trans. Multimdeia*. (to appear)

[8] Y.K.Yu, K.H.Wong, M.Y.Y.Chang and S.H.Or, "Recursive camera motion estimation with the trifocal tensor", *IEEE Trans. Syst., Man, Cybern. B*. (to appear)

[9] A.J.Davison and N.Kita, "3D simultaneous localisation and map-building using active vision for a robot moving on undulating terrain", presented at IEEE Conf. Comput. Vision Pattern Recognit., Kauai, Dec. 2001.

[10] 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. IEEE Int. Conf. Robotics Autom.*, pp. 414-419, Ohio, May 1990.

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

[12] C.Tomasi and T.Kanade, "Detection and tracking of point features", Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-CS-91-132, Apr. 1991.

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

[14] B.Triggs, P.McLauchlan, R.Hartley and A.Fitzgibbon, "Bundle adjustment – A modern synthesis", in *Proc. Intl. Workshop Visual Algorithm: Theory and Practice*, pp. 298-372, Corfu Greece, 1999.

[15] C.J.Poelman and T.Kanade, "A paraperspective factorization method for shape and motion recovery", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 3, pp. 206-218, Mar. 1997.

[16] M.M.Y.Chang and K.H.Wong, "Model reconstruction and pose acquisition using extended Lowe's method", *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 253-260, Apr. 2005.

[17] S.Avidan and A.Shashua, "Threading fundamental matrices", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 1, pp. 73-77, Jan. 2001.

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

[19] S.Hutchinson, G.D.Hager and P.I.Corke, "A tutorial on visual servo control", *IEEE Trans. Robotics Autom.*, vol. 12, no. 5, pp. 651-670, Oct. 1996.

[20] Y.K.Yu, K.H.Wong and M.M.Y.Chang, "Pose estimation for augmented reality applications using genetic algorithm", *IEEE Trans. Syst., Man, Cybern. B*, vol. 35, no. 6, pp. 1295-1301, Dec. 2005

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

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

[23] J-Y.Bouguet, Camera Calibration Toolbox for Matlab. (http://www.vision.caltech.edu/bouguetj/calib_doc/)

[24] B.Lloyd, Computation of the Fundamental Matrix. (http://www.cs.unc.edu/~blloyd/comp290-089/fmatrix/)

[25] D.Nister, "Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors", presented at European Conf. Comput. Vision, Ireland, Jun. 2000.

[26] S.H.Or, W.S.Luk, K.H.Wong and I.King, "An efficient iterative pose estimation algorithm", *Image Vision Comput.*, vol. 16, no. 5, pp. 355-364, Apr. 1998.