COMPUTING POSE SEQUENCES DIRECTLY FROM VIDEOS

Ying Kin Yu[1], Kin Hong Wong[1], Michael Ming Yuen Chang[2] and Siu Hang Or[1]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.
[2] Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong.
Email: {ykyu, khwong, shor}@cse.cuhk.edu.hk, mchang@ie.cuhk.edu.hk

ABSTRACT

An innovative extended Kalman filter (EKF) algorithm for pose tracking has been proposed in this paper. It has the advantages of both structure and motion-based (SAM-based) and traditional model-based pose estimation algorithms. With no prior information about the scene, the pose sequence can be computed directly from images while the updating of the 3-D structure is not necessary. To achieve the goal, a constant velocity motion model is used as the dynamic system and the trifocal tensor point transfer function is applied to the measurement model of the filter. The resulting algorithm is stable, accurate and efficient. An empirical comparison with existing EKFs which deal with the same problem has been made and shows that our approach outperformed the others. The proposed method has been tested with various video sequences to demonstrate its performance in real situations.

Index Terms— Image motion analysis, machine vision, Kalman filtering.

1. INTRODUCTION

A fast and robust pose acquisition algorithm is crucial to interactive applications such as augmented reality and robot navigation. An accurate pose estimation method is also important for the recovery of the 3-D structure, since a high precision depth map can be constructed with an optimal pose sequence.

Traditional pose acquisition methods require known 3-D structure of the scene [15] [16]. More general approaches, which require no prior information on the 3-D model, are based on the techniques in structure and motion (SAM) such as multiple view geometry [10] [11], factorization [1] and bundle adjustment [3] [13]. To deal with the problem recursively, Kalman filtering can be applied [4] [5] [6] [7] [8] [9] [17] [18]. The work in [6], which uses an iterated extended Kalman filter to achieve the task, is the seminal work. Azarbayejani and Pentland [5] proposed an improvement of [6] by making an extension in recovering the camera focal length and the representation of the 3-D model. In [8], Yu *et. al.* decoupled the full covariance extended Kalman filter (EKF) such that the computation efficiency is increased as a tradeoff in accuracy. An extension of their work can be found in [9], in which the Interacting Multiple Model (IMM) was added to the original formulation. Soatto *et. al.* [12] applied the essential constraint to pose estimation but this constraint is susceptible to degeneracy in some real situations [10].

This paper describes an innovative EKF-based algorithm that tackles the pose tracking problem. With no prior information about the 3-D structure of the scene, the pose information can be recovered from a monocular image sequence directly and recursively with the trifocal tensor. The major contribution of our approach is the incorporation of the trifocal tensor into the Kalman filtering formulation. In the algorithm, the trifocal tensor point transfer function is used in the measurement model of the EKF. That is apart from the dynamic system constraint on the motion of the camera, the trifocal constraint has also been employed in the filtering cycle, resulting in an enhancement on the accuracy of the algorithm.

In our formulation, the recovery of pose sequence is independent of the 3-D structure. Traditional recursive SAM-based algorithms require the update of both the 3-D model points and pose parameters either simultaneously [5] or in an interleaved manner [8]. As no computation of the 3-D model coordinates are involved in our approach, the total number of parameters required to be estimated is reduced from N+6, where N is number of available point features, to 12. In this way, both the stability and computation efficiency of our filter can be improved.

An empirical comparison with existing recursive SAM-based pose tracking approaches [5] [8] has been made. Experimental results show that our algorithm had a better overall performance than the others. In addition, the proposed approach has been tested using real video sequences. Its accuracy has been verified by re-projecting the corner features in the 1st image to the succeeding frames with the recovered pose.

2. PROBLEM MODELING

The relationship between a point X_m^O of the 3-D structure in the world coordinate frame and its 2-D projection $p_{m,t}$ on the image plane can be related as:

$$X_{m,t}^C = R_t X_m^O + T_t \quad (1)$$

$$p_{m,t} = \begin{bmatrix} u_{m,t} \\ v_{m,t} \end{bmatrix} = \frac{f'}{z_{m,t}^C} \begin{bmatrix} x_{m,t}^C \\ y_{m,t}^C \end{bmatrix} \quad (2)$$

where $X_{m,t}^C = [x_{m,t}^C, y_{m,t}^C, z_{m,t}^C]^T$ denotes the coordinates of X_m^O in the camera coordinate frame and the subscript t denotes the time when the measurement is being made. f' represents the focal length of the camera. $T_t = [x_t \quad y_t \quad z_t]^T$ is a 3×1 translation vector and R_t is a 3×3 rotation matrix parameterized by the Yaw (α_t), Pitch (β_t) and Roll (γ_t) angle. The objective of the proposed algorithm is to recover the pose information R_t and T_t at each time-step recursively given only the image measurements $p_{m,t}$.

3. SUMMARY OF THE ALGORITHM

Fig. 1 shows the overview of the proposed pose tracking algorithm. The Kanade-Lucas-Tomasi (KLT) tracker described in [2] is used to extract feature points and track them in the images. It is assumed that the point features extracted by the tracker are contaminated only by Gaussian noise.

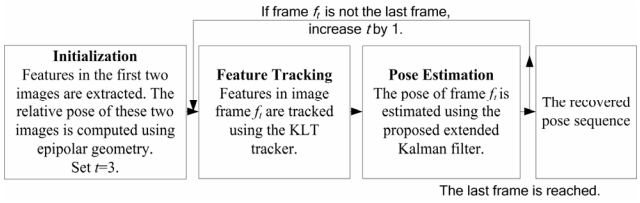


Fig. 1. The flowchart of the proposed pose tracking algorithm.

The algorithm is initialized by estimating the relative pose of the first two image frames f_1 and f_2 using epipolar geometry. Specifically, the fundamental matrix is first computed using the 8-point algorithm plus a RANSAC robust estimator. The pose parameters R_2 and T_2, which is up to an unknown scale factor, are then extracted from the fundamental matrix [10]. This is actually an initial guess of the pose of image f_2.

Starting from image f_3, the measurements are processed by an EKF. In each cycle, three images are processed, within which two of them are images f_1 and f_2 in the sequence. They compose of the base frames of the filter. The third one is the image f_t at the current time-step t. The EKF computes the pose of image f_t and, at the same time, refines the initial guess of the pose of image f_2. Its state vector, denoted by w_t, is written as:

$$w_t = [\ x_t \quad \dot{x}_t \quad y_t \quad \dot{y}_t \quad z_t \quad \dot{z}_t \quad \alpha_t \quad \dot{\alpha}_t \quad ... \quad \beta_t \quad \dot{\beta}_t \quad \gamma_t \quad \dot{\gamma}_t \quad x_2 \quad y_2 \quad z_2 \quad \alpha_2 \quad \beta_2 \quad \gamma_2 \ ]^T \quad (3)$$

$\dot{x}_t, \dot{y}_t, \dot{z}_t$ are the translational velocities corresponding to x_t, y_t, z_t while $\dot{\alpha}_t, \dot{\beta}_t, \dot{\gamma}_t$ are the angular velocities corresponding to $\alpha_t, \beta_t, \gamma_t$. The state transition and measurement equation of the filter are formulated as:

$$w_t = A w_{t-1} + \gamma'_t \quad (4)$$

$$\varepsilon'_t = g_t(w_t) + v'_t \quad (5)$$

where

$$g_t(w_t) = [u_{1,t} \quad v_{1,t} \quad ... \quad u_{m,t} \quad v_{m,t} \quad ... \quad u_{N,t} \quad v_{N,t}] \quad (6)$$

$$[U_{m,t}]^k = [U_{m,1}]^i [l_{m,2}]_j [\Gamma_t]_i^{jk} \quad (7)$$

and

$$A = diag\left\{ \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} ... \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & S \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

γ'_t and v'_t are zero mean Gaussian noise. The dynamic system models the camera as constant velocity motion with the initial guess of the pose R_2 and T_2 close to the actual values. ε'_t is an $N \times 1$ column vector representing the measurements from images, where N denotes the number of point features input to the filter. $g_t(w_t)$ is the $N \times 1$-output trifocal tensor point transfer function. Equation (7) is written in the tensor notation, in which Γ_t is known as the trifocal tensor [10]. Γ_t encapsulates the geometric relations among three views and contains the pose parameters R_2, T_2, R_t and T_t. The use of trifocal tensor here makes the recovery of pose information directly from 2-D images possible. $U_{m,t}$ is the normalized homogenous form of $p_{m,t}$ such that $U_{m,t} = [\bar{u}_{m,t} \quad \bar{v}_{m,t} \quad \bar{w}_{m,t}]^T = [u_{m,t}/f' \quad v_{m,t}/f' \quad 1]^T$. $l_{m,2}$ is a line passing through the point $p_{m,2}$ in image f_2. S represents the sample period of the image measurements. The implementation details of the EKF and the construction of tensor Γ_t plus line $l_{m,2}$ can be found in [14] and [10], respectively. The filtering loop ends when all the images are processed.

If the set of available point features is changing in the image sequence, the additional procedure to handle the case is to find the set of features commonly appeared in all the three images being processed in a filtering cycle. If the set of available features, say extracted from f_1, f_2 and f_t, processed by the filter falls below 7, the algorithm is bootstrapped. The process followed is that images f_{t-o} and f_{t-o+1} are used to re-initialize the algorithm and become the base frames. A new filtering loop is then started from image f_{t-o+2}. Here o is the number of image frames to be re-computed in the new filtering loop. For example, o is set to 5 in our implementation. In this way, the scale of the translation parameters (before and after re-initialization) can be aligned.

4. EXPERIMENTS AND RESULTS

4.1. Experiments with synthetic data

An object with 300 random feature points in 3-D 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 $\alpha_t, \beta_t, \gamma_t$ and 0.005 to 0.015 meters per frame for x_t, y_t and z_t. The length of each synthetic sequence was 99 frames. A total of 50 independent tests were carried out. The proposed algorithm, the EKF by Azarbayejani and Pentland [5] and the 2-step EKF by Yu *et. al.* [8] were implemented in Matlab and run on a Pentium III 1GHz machine to estimate the pose information. The results were compared and analyzed.

Fig. 2 shows the average accumulated total rotation and translation errors of the three approaches. The line with asterisk (*), triangle (△) and circle (○) markers are for our approach, the EKF by Azarbayejani and Pentland [5] and the 2-step EKF by Yu *et. al.* [8], respectively. Here, the total rotation of the camera were calculated using the axis-angle representation, with which the Yaw, Pitch, Roll angle was reduced into a single angle. The difference between the actual and the recovered value is the accumulated error. The accumulated total translation error was computed by subtracting the recovered translation vector from the actual one and the magnitude was taken. From the plots, it is clear that the proposed approach had a lower error than the other methods.

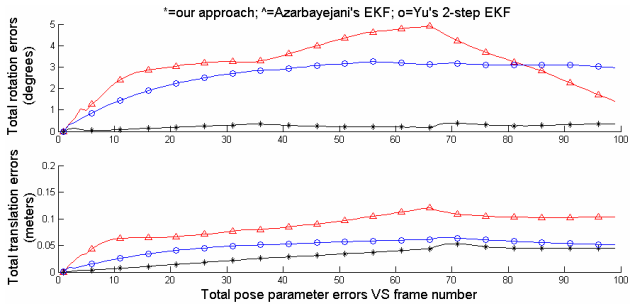


Fig. 2. The average accumulated total rotation error (top, in degrees) and accumulated total translation error (bottom, in meters) versus frame number of the 3 algorithms.

TABLE I
TIME REQUIRED TO PROCESS AN EXTRA IMAGE

	Our approach	Azarbayejani's EKF	Yu's 2-step EKF
Time required (seconds)	1.56	2.60	0.42

A table showing the average CPU time for the 3 algorithms to recover the pose when extra frames were added to the image sequence.

Table I shows the time needed to recover the pose when new image measurements were sequentially fed to the algorithms. Our algorithm outperformed the full covariance EKF by Azarbayejani and Pentland. However, the 2-step EKF took the shortest time to achieve the task. The reason is that their EKF is decoupled, which is actually a tradeoff between speed and accuracy.

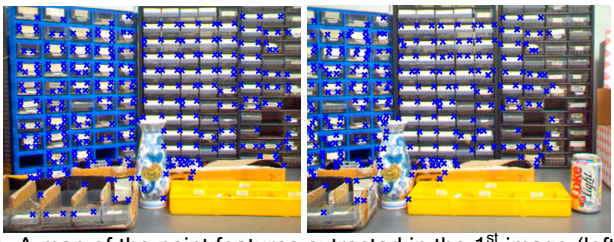
4.2. Experiments with real images

Two real image sequences were used to test the proposed approach. The first sequence was taken in the laboratory. The images were captured while the camera was translating sideway on a rig. The length of the image sequence is 100 frames. In the second sequence, the Grand Canyon in northwestern Arizona was viewed from an airplane. It is 5-second long and contains 50 images. The proposed algorithm was applied to track the changes of the pose.

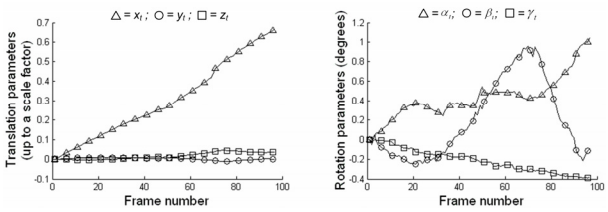
Figs. 3 and 4 show the results. To verify whether the recovered pose is correct, corner features in the 1st image of the sequence were extracted. A set of trifocal tensors was computed using the pose parameters recovered. It was used to transfer (re-project) the corner features from the 1st image to the succeeding frames. We checked the consistency of the motion of these corner features with respect to the background images. From Figs. 3a and 4a, you can see that the features, which are indicated by cross markers, stick to the same position relative to the background. We can say that the results are accurate and visually acceptable. Figs. 3b and 4b illustrate the values of the acquired pose parameters. The line with triangle (△), circle (○) and square (□) markers on the left plot are respectively for the translation parameters x_t, y_t and z_t while the line with triangle (△), circle (○) and square (□) markers on the right plot are respectively for α_t (Yaw), β_t (Pitch) and γ_t (Roll) angle. By inspecting the motion in the original videos, the recovered pose sequences are reasonable. More results can be found at http://www.cse.cuhk.edu.hk/~khwong/demo/

5. CONCLUSION

A high-speed recursive pose tracking algorithm has been proposed in this paper. By integrating the trifocal tensor with the extended Kalman filter, a significance improvement on the accuracy and computation efficiency has been achieved. The pose sequence can be recovered directly from images without the explicit reconstruction of 3-D structure. Thus, the procedure to handle the changeable set of point features becomes simple. It is shown in the experiment that the proposed algorithm is accurate in both simulations and real situations. Our novel approach is suitable for a wide range of image processing applications such as augmented reality and visual servoing.

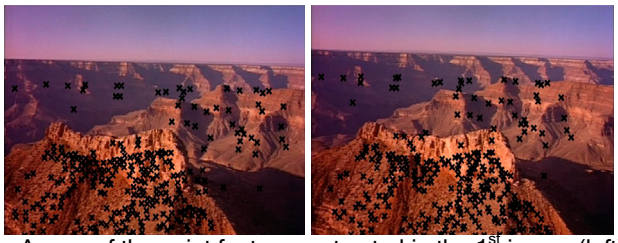


a. A map of the point features extracted in the 1st image (left) and its re-projection on the 50th image (right).

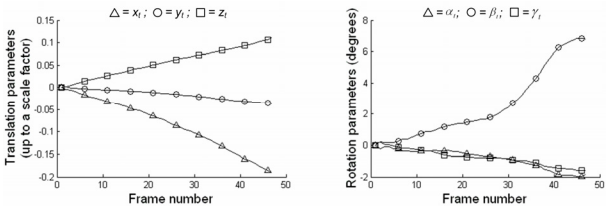


b. The recovered pose parameters.

Fig. 3. Results of the laboratory scene sequence.



a. A map of the point features extracted in the 1st image (left) and its re-projection on the 50th image (right).



b. The recovered pose parameters.

Fig. 4. Results of the Grand Canyon sequence.

6. 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.

7. REFERENCES

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, *Cybern.*, vol. 35, no. 3, pp. 587-592, Jun. 2005.

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

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

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

[12] S.Soatto, R.Frezza and P.Perona, "Motion Estimation on the Essential Manifold", presented at ECCV, Stockholm, Sweden, May 1994.

[13] Z.Zhang and Y.Shan, "Incremental motion estimation through modified bundle adjustment", in *Proc. IEEE ICIP*, vol. 2, pp. 343-346, Barcelona, Spain, Sep. 2003.

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

[15] D.G.Lowe, "Fitting parameterized three-dimensional models to images", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 5, pp. 441-450, May 1991.

[16] 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*, *Cybern.*, vol. 35, no. 6, pp. 1295-1301, Dec. 2005.

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

[18] 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. IEEE CVPR*, New York, Jun. 2006.