International Conference on Image Processing (ICPR04), Singapore, Oct. 2004.

# A FAST AND ROBUST SIMULTANEOUS POSE TRACKING AND STRUCTURE RECOVERY ALGORITHM FOR AUGMENTED REALITY APPLICATIONS

*Ying-Kin Yu[1], Kin-Hong Wong[1] and Michael Ming-Yuen Chang[2]*

[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}@cse.cuhk.edu.hk, mchang@ie.cuhk.edu.hk

## ABSTRACT

*A robust simultaneous pose tracking and structure recovery algorithm based on the Interacting Multiple Model (IMM) for augmented reality applications is proposed in this paper. A set of three extended Kalman filters (EKFs), each describes a frequently occurring camera motion in real situations (general, pure translation, pure rotation), is applied within the IMM framework to track the pose of an object. Another set of EKFs, one filter for each model point, is used to refine the positions of the model features in the 3D space. The filters for pose tracking and structure refinement are executed in an interleaved manner. The results are used for inserting virtual objects into the original video footage. The performance of the algorithm is demonstrated with both synthetic and real data. Comparisons with different approaches have been performed and show that our method is more efficient and accurate.*

## 1. INTRODUCTION

The research work presented in this paper falls into the category of structure and motion (SAM) in computer vision. The goal is to reconstruct a 3D structure and its pose from a sequence of 2D images. A major stream of the solutions is to tackle the problem in a batch. Factorization [2] and bundle adjustment [4] are common approaches. Besides, there are solutions that deal with the problem in a recursive way. Most of them are based on Kalman filtering. The work in [1] adopts iterated extended Kalman filter (IEKF) for structure updating. The series of methods in [5] [6] [7] recover both the structure and motion. The work in [7] is the ancestor of this series of researches. The authors apply a single IEKF to recover the structure and pose of an object. Azarbayejani and Pentland describe a method in [6] that improves [7] by making an extension in recovering the camera focal length and the representation of the 3D model. The most recent recursive method is by Chiuso et al [5]. Similar Kalman filtering techniques in SAM have also been applied to simultaneous localization and map-building for robot navigation [8].

The Interacting Multiple Model (IMM) based algorithm presented in this paper aims to solve the problem of structure and pose ambiguities encountered in most of the existing SAM algorithms, which has been reported by Szeliski and Kang in [9]. Inaccuracy due to the ambiguities can be minimized if prior information about the structure or motion is utilized. In our algorithm, the three extended Kalman filters (EKFs) for pose estimation embedded within the IMM framework describe three different motion dynamics that represent frequently occurring camera motions in real situations. The IMM [10] provides a mechanism to "select" suitable filters automatically in order to set constraints on the camera's motion once prior information is available. With the constraints, the total number of parameters to be estimated is reduced and the accuracy can be improved. In addition, the problem of motion discontinuity in real image sequences can be handled properly with the IMM.

Our structure and motion algorithm consists of $N$+3 small EKFs. This arrangement has the advantage of avoiding tripling the computation time. Indeed, it reduces the time complexity from quadratic to linear compared to the approaches that use a single full covariance EKF. It is necessary since the computation speed is crucial for augmented reality applications. The rigidity of the object under reconstruction is maintained by expressing the 3D point features in terms of the first images that they appear. Experimental results show that our IMM-based approach can resolve the ambiguities among the Yaw angle, Pitch angle and the structure notably. Our method also has higher computation efficiency than the interleaved bundle adjustment method [4] and the EKF by Azarbayejani and Pentland [6]. We have tested our algorithm by tracking the pose of a real scene. The recovered pose sequence has been applied to produce an augmented reality video.

## 2. PROBLEM MODELING

Figure 1 shows the geometry of our system. $X_i^O = [x_i^O, y_i^O, z_i^O]^T$ and $X_{i,t}^C = [x_{i,t}^C, y_{i,t}^C, z_{i,t}^C]^T$ denote the coordinates of the point $X_i$ with respect to the object and the camera coordinate frame respectively. $p_{i,t} = [u_{i,t}, v_{i,t}]^T$

is a point on the image plane. The recovered structure is centered at the origin $O_o$. The relationship between the object frame and the camera frame is as follows:

$$X_{i,t}^{C} = (R_t X_i^{O} + T_t) + T^{C} \quad (1)$$

$R_t$ is a 3x3 rotation matrix and $T_t$ is a 3x1 translation vector. $T^C$ is a 3x1 vector that brings the object in the object frame to the camera frame. Parameter $R_t$ and $T_t$ compose of the pose sequence. The camera is calibrated with fixed focal length $f$. The camera model is full perspective and the projection can be expressed as:

$$\begin{bmatrix} u_{i,t} \\ v_{i,t} \end{bmatrix} = \frac{f}{z_{i,t}^{C}} \begin{bmatrix} x_{i,t}^{C} \\ y_{i,t}^{C} \end{bmatrix} \quad (2)$$

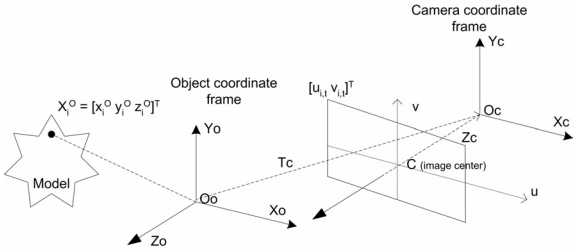

Figure 1. The geometry of our system.

## 3. OVERVIEW OF THE ALGORITHM

The system can be divided into four parts: feature extraction and tracking, model initialization, pose estimation and structure updating.

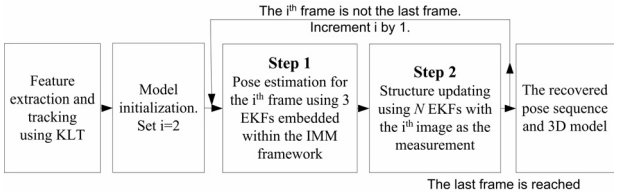

Figure 2. The flowchart of our IMM-based SAM algorithm.

The KLT tracker described in [3] is used to extract feature points and track them in the images. The 3D model is initialized by assuming that the projection of the first image in the sequence is orthographic.

The initial model and the second image are fed to the first step of the main loop for pose estimation. Three extended Kalman filters (EKFs), each represents a unique motion dynamics, are adopted. These three filters interact with one another using the Interacting Multiple Model (IMM) [10]. The recovered pose and the input image are passed to the second step for structure updating.

The second step consists of a set of $N$ EKFs. Each EKF corresponds to each coordinate point in the recovered 3D structure. With the observations and the pose recovered for the current image, the coordinates of each feature point are updated accordingly. The algorithm alternates between the step 1 and 2 until all images in the sequence are used.

## 4. STEP 1: POSE ESTIMATION

The three EKFs describing three different motion dynamics are defined as follows:

1) The General Motion Filter (GMF): GMF is designed to handle arbitrary object motion with unrestricted rotation and translation. Constant velocity is assumed for the GMF.
2) The Pure Translation Motion Filter (TMF): TMF is designed for tracking the objects with zero rotation motion.
3) The Pure Rotation Motion Filter (RMF): RMF is dedicated to tackle the objects with pure rotation around the y-axis (i.e. non-zero pitch angle).

The state vector $w$ is common for all the filters:

$$w = \begin{bmatrix} t_x & \dot{t}_x & t_y & \dot{t}_y & t_z & \dot{t}_z & \alpha & \dot{\alpha} & \beta & \dot{\beta} & \gamma & \dot{\gamma} \end{bmatrix}$$

$t_x$, $t_y$, and $t_z$ are the translation parameters of the object along the x, y and z axis respectively. $\dot{t}_x, \dot{t}_y, \dot{t}_z$ are their corresponding velocities. $\alpha, \beta, \gamma$ are respectively the Yaw, Pitch and Roll angle with $\dot{\alpha}, \dot{\beta}, \dot{\gamma}$ as their corresponding angular velocities. The state transition and measurement equation for the filters are:

$$w_t = A w_{t-1} + \gamma'_t$$
$$\varepsilon'_t = g_t(w_t) + \nu'_t$$

$\gamma'_t$ and $\nu'_t$ are zero mean Gaussian noise. $A$ is a 12x12 block diagonal state transition matrix. $A$ is different for the three motion filters. $\varepsilon'_t$ is an $n$x1 column vector representing the selected real measurements from the images. $g_t(w)$ is the $n$x1-output projection function similar to equation (2). With the above equations, the EKF implementation is straightforward and can be found in related textbooks.

In brief, the IMM algorithm consists of four steps. Firstly, the filter likelihood, states and covariances of the motion filters are updated according to the switching matrix. These values are used by the EKFs defined previously in a usual way. After the EKF cycle, the new filter likelihood is computed according to the image residuals and residual covariances of the corresponding filters. Lastly, the usable output state and covariance are generated with the smoothed states and covariances weighted by the updated filter likelihood. The pose estimation step is finished.

## 5. STEP 2: STRUCTURE UPDATING

For $N$ model points, $N$ EKFs are needed for the structure update. The model is assumed to be static. The dynamic model of a 3D point and the measurement equation are:

$$X'_{i,t} = X'_{i,t-1} + \gamma_t$$
$$\varepsilon_{i,t} = h_t(X'_i) + v_t$$

$\gamma_t$ and $v_t$ are the zero mean Gaussian noise. $\varepsilon_{i,t}$ represents the real measurements from the image sequence. $h_t(X'_i)$ is the projection function, which can be found by substituting $X'_i$ into equation (3) (1) and then (2). $X'_i$ is a scalar that represents a model point:

$$X_i^O + T^C = \begin{bmatrix} x_i^S \\ y_i^S \\ z_i^S \end{bmatrix} = \begin{bmatrix} u_{i,1} \\ v_{i,1} \\ 0 \end{bmatrix} + \frac{X'_i}{f} \begin{bmatrix} u_{i,1} \\ v_{i,1} \\ f \end{bmatrix} \quad (3)$$

Intuitively, the 3D coordinates of the points are expressed in terms of the first images that the features appear. This measure reduces the required computation time and maintains the rigidity of the object under the assumption that the measurements acquired are non-biased. Again, the implementation of EKF in this step is standard and the formulation is not repeated here.

## 6. EXPERIMENTS AND RESULTS

### 6.1. Experiments with synthetic data

An object with 300 random feature points in 3D within a cube of volume of 0.13m$^3$, centered at a place 0.33m away from the camera, was generated. The motion of the object was composed of three different segments, a pure translation section, a pure rotation section and a general motion section. The sequence of occurrence of the sections was by random. The motion parameters were generated randomly from 0.05 to 0.15 degrees per frame for Yaw, Pitch, Row angle and 0.0005 to 0.0015 meters per frame for $t_x$, $t_y$ and $t_z$. The length of each synthetic sequence is 60 frames. A total of ten independent tests were carried out. Our IMM-based algorithm, the interleaved bundle adjustment method [4] and the EKF by Azarbayejani and Pentland [6] were tested and compared.

Figure 3 shows the average total rotation and translation errors of the three approaches under the ten test cases. For the plots in figure 3 to 5, the line with asterisk (*), triangle (¢) and circle (O) markers are for our IMM-based approach, the interleaved bundle adjustment method and the EKF by Azarbayejani and Pentland respectively. These three algorithms were implemented in Matlab with a Pentium III 1GHz machine and the time measurement is in seconds. In figure 3, the total rotation error is calculated by using the axis-angle representation to reduce the Yaw, Pitch, Roll angle into a single angle. The total translation error is computed by Pythagoras theorem. Our approach has errors lower than the other two methods most of the time. It is obvious that our approach can recover a more accurate pose.

Figure 4 shows the time for the three algorithms to optimize the image residual error of the back-projected model. Our algorithm falls to the lowest errors at the earliest time among the three methods. It can resolve the structure and pose ambiguities to at least a certain extent and thus avoids the poor local optima.

Figure 5 shows the time needed to reconstruct a model when extra frames were added sequentially to the image sequence. The first step in creating this plot was to reconstruct a model with the first 10 frames. The succeeding 50 frames were sequentially fed to the algorithm as the new measurements of the scene. You can see that our approach outperformed the other two algorithms. Our algorithm takes only 0.79 seconds to update the structure of the scene for every extra frame added to the image sequence. The EKF by Azarbayejani and Pentland takes 2.60 seconds while the interleaved bundle adjustment method takes at least 4.55 seconds.

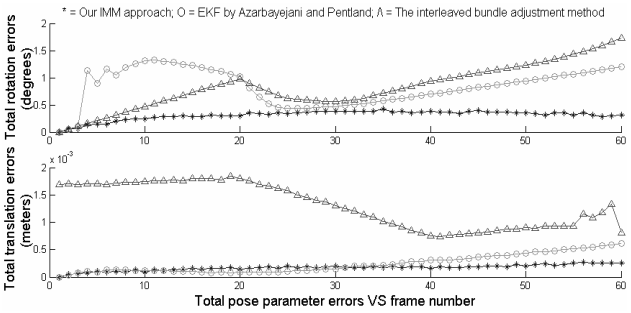


Figure 3. The average total rotation error (top, in degrees) and total translation error (bottom, in meters) versus frame number of the 3 algorithms.

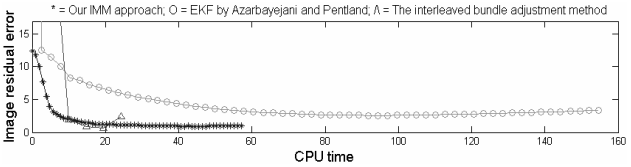


Figure 4. The relationship between the CPU time and the image residual error.

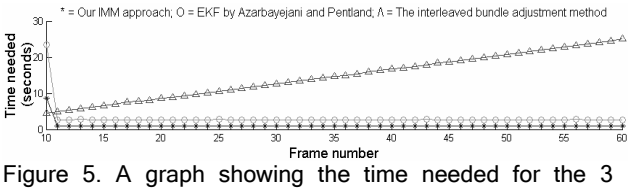


Figure 5. A graph showing the time needed for the 3 algorithms to reconstruct the model and pose when extra frames are added to the image sequence.

### 6.2. Experiments with real scene

An experiment using real scene images was also performed. The test image sequence was captured by translating the camera sideway on a rig. The length of the image sequence is 100 frames. Our IMM-based algorithm was applied to track the camera motion while reconstructing the scene's structure. The recovered pose

sequence was used to produce an augmented reality video, in which a synthetic car was put onto the yellow box in the real scene.

Figure 6. Results of inserting an artificial object into the real scene. First row: The first and the last image of the laboratory sequence. Second row: A synthetic car, which is drawn by wire-frames, was inserted into the real scene. Demonstration video can be found at http://www.cse.cuhk.edu.hk/~khwong/demo/

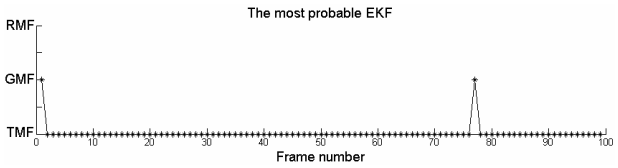


Figure 7. A plot showing the most probable EKF for pose estimation against frame index in the process of tracking the camera motion of the real image sequence.

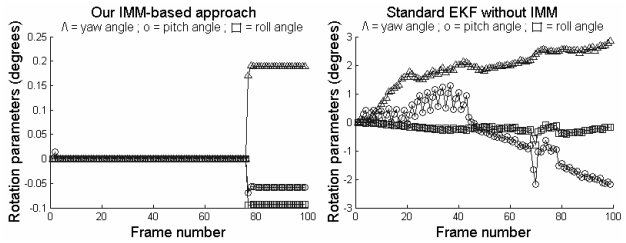


Figure 8. The rotation parameters recovered from the real scene with our IMM-based approach (the left one) and with a standard EKF with no IMM (the right one).

Figure 6 shows the results. An augmented reality video has been made successfully. The orientation of the synthetic car is consistent with the real scene. Figure 8 shows a comparison of the rotation parameters acquired with and without the IMM. The pose recovered with our IMM-based algorithm is smooth with no ambiguities among the Yaw and Pitch angle. The small jump in the rotation angles at the 76$^{th}$ frame is due to the vibration of the camera. On the other hand, the pose recovered using a standard EKF with no IMM (i.e. A 2-step EKF in which one GMF is for pose estimation and $N$ EKFs are for structure updating) is fluctuating and has an error about 2 degrees for the Yaw and Pitch angle.

## 7. CONCLUSION

The ambiguities between pose and structure have been solved by our IMM-based structure and motion algorithm under the following frequently occurring camera motions: pure translation, pure rotation or motion discontinuity. Our approach outperformed two other existing SAM algorithms in both the accuracy of the recovered pose and computation speed. Our algorithm also achieves a linear time complexity in terms of the total available point features, which suits the real-time requirement for many of the augmented reality applications.

## 8. REFERENCES


[1] P.A. Beardsley, A.Zisserman and D.W.Murray, "Sequential updating of projective and affine structure from motion", IJCV vol. 23, no. 3, pp. 235-259, 1997.

[2] C.Tomasi and T.Kanade, "Shape and motion from image streams under orthography: A factorization method", IJCV vol. 9, no. 2, pp. 137-154, 1992.

[3] C.Tomasi and T.Kanade, "Detection and Tracking of Point Features", Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

[4] B.Triggs, P.McLauchlan, R.Hartleyand and A.Fitzgibbonm, "Bundle adjustment – A modern synthesis" In proc. of the Intl. Workshop on Visual Algorithm: Theory and Practice, pp. 298-372, Corfu Greece, 1999.

[5] A.Chiuso, P.Favaro, H.Jin and S.Soatto, "Structure from motion casually integrated over time", IEEE Trans. on PAMI, vol. 24, no. 4, 2002.

[6] A.Azarbayejani and A.P.Pentland, "Recursive estimation of motion, structure, and focal length", IEEE Trans. on PAMI, vol. 17, no. 6, June 1995.

[7] T.J.Broida, S.Chandrasiiekhar and R.Chellappa, "Recursive 3-D motion estimation from monocular image sequence", IEEE Trans. on Aerospace and Electronic Systems, vol. 26, no. 4, July 1990.

[8] A.J.Davison and D.W.Murray, "Simultaneous localization and map-building using active vision", IEEE Trans. on PAMI, vol. 24, no. 7, July 2002.

[9] R.Szeliski and S.B.Kang, "Shape ambiguities in structure from motion", IEEE Trans. on PAMI, vol. 19, no. 5, May 1997.

[10] E.Mazor, A.Averbuch, Y.Bar-Shalom and J.Dayan, "Interacting multiple model methods in target trackings: A Survey", IEEE Trans. on Aerospace and Electrons Systems, vol. 34, no. 1, pp. 103-123, January 1998.