In Proc. of The IEEE 2006 International Conference of Pattern Recognition (ICPR2006), Hong Kong, Aug. 2006.

# Accurate 3-D Motion Tracking with an Application to Super-Resolution

Ying Kin Yu[1], Siu Hang Or[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 , shor, khwong}@cse.cuhk.edu.hk, mchang@ie.cuhk.edu.hk*

## Abstract

*Many of the existing image processing applications, in particular the construction of super-resolution videos, require an accurate high-speed motion tracking algorithm. This paper proposes an efficient recursive approach that recovers 3-D motion from a stereo image sequence with high precision based on the trifocal tensor. In the computation process, neither the 3-D structure of the scene nor its reconstruction is necessary. The validity of the proposed algorithm is demonstrated by applying it to upgrade the resolution of real stereo images. Empirical comparisons show that our novel approach outperformed traditional work in terms of both the accuracy of motion estimation and the quality of super-resolved images.*

## 1. Introduction

Traditional 3-D motion estimation methods assume that the model structure is known in advance [6] [18]. More general approaches are built with the techniques in the structure and motion (SAM) algorithms [9] [10], which recover both the 3-D structure and camera motion simultaneously from 2-D images. SAM-based motion tracking algorithms with high-speed and low latency rely on the use of Kalman filters. The series of methods in [1] [2] [3] [4] [5] recover the structure and motion from a monocular image sequence. In the extended Kalman filter (EKF)-based algorithm developed by Azarbayejani and Pentland [5], the focal length of the camera can be estimated and the pointwise structure is represented by one parameter. Yu *et. al.* [1] decoupled the traditional full covariance EKF such that the computation efficiency is increased as a tradeoff in accuracy. Soatto *et. al.* [11] applied the essential constraint in epipolar geometry to recursive motion estimation. However, the essential matrix becomes degenerate under some commonly appeared motions in real-life [9].

In this paper, we propose a stereo 3-D motion tracking method that is applicable to the super-resolution problem. Super-resolution refers to the construction of a high resolution image from a number of lower resolution originals of the same scene [13] [14] [15]. The major advantage of our approach over existing SAM-based methods is that neither known 3-D structure nor its computation is required while recovering the motion. As the handling of 3-D models is not necessary, the values that are computed in each computation cycle are limited to the six velocity parameters of the 3-D motion only, leading to an improvement on the estimation accuracy and computation efficiency.

## 2. The 3-D motion tracking problem

The relationship between a 3-D point $X_m^W = [x_m^W \quad y_m^W \quad z_m^W]^T$ in the world coordinate frame and its projection on the left image $p_{m,t}$ and right image $p'_{m,t}$ of the stereo system are expressed as:

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

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

where $K$ is a $3\times3$ matrix that encodes the intrinsic parameters, say focal length $f$, of the camera. $E$ is a $3\times4$ matrix representing the rigid transformation between the two cameras and is assumed fixed. $M_t$ is a $4\times4$ matrix that transforms the 3-D structure of the scene from the world frame to the reference (left) camera frame at time instance *t*. The objective of the proposed pose tracking algorithm is to compute the 3-D camera motion, i.e. $M_t$, at each time-step recursively given only the image measurements $p_{m,t}$ and $p'_{m,t}$.

## 3. The proposed motion estimation algorithm

Fig. 1 summarizes the proposed 3-D motion tracking algorithm. The Kanade-Lucas-Tomasi (KLT) tracker described in [8] is used to extract feature points and track them in the images.

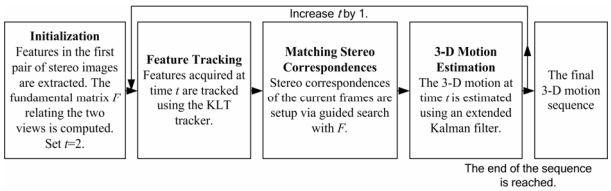


Fig. 1. A summary of our 3-D motion tracking algorithm.

The features from the left and the right image sequences are tracked independently. Their stereo correspondences are setup afterwards using the procedure in [12]. As the fundamental matrix $F$ has also been computed in the process, the required extrinsic parameters $E$ of the stereo system can be extracted according to [9] with known intrinsic parameters $K$.

The EKF, in company with the trifocal tensor, are then used to estimate the camera motion from a stereo image sequence. The state vector $\dot{\xi}_t$ of the EKF is defined as:

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

$\dot{x}_t, \dot{y}_t, \dot{z}_t$ are the translational velocities of camera along the $x$, $y$ and $z$ axis, respectively. $\dot{\alpha}_t, \dot{\beta}_t, \dot{\gamma}_t$ are respectively the angular velocities of camera rotation on the $x$, $y$ and $z$ axis.

The EKF models the dynamics of the camera as acceleration having a zero-mean Gaussian value $\eta_t$. Assuming that the sampling rate of the measurements is high and the camera motion between the successive images in a sequence is small, the dynamic system equations of the filter are as follows:

$$\begin{aligned} M_t &= M_{t-1} \exp(\dot{\xi}_t) = M_{t-1}(I + \tilde{\dot{\xi}}_t) \\ \dot{\xi}_t &= \dot{\xi}_{t-1} + \eta_t \end{aligned} \quad (4)$$

where $\tilde{\dot{\xi}}_t$ is the matrix form of the twist $\dot{\xi}_t$ [17]. The measurement model, which relates the 3-D motion $M_t$ and the measurements $\varepsilon_t$ acquired from the pair of stereo cameras, is defined as:

$$\varepsilon_t = g_t(M_t) + v_t \quad (5)$$

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 available point features. $g_t(M_t)$ is the $4N \times 1$-output trifocal tensor point transfer function and is expressed as:

$$\begin{aligned} g_t(M_t) = \big[ & u_{1,t} \quad v_{1,t} \quad \ldots \quad u_{m,t} \quad v_{m,t} \quad \ldots \\ & u_{N,t} \quad v_{N,t} \quad u'_{1,t} \quad v'_{1,t} \quad \ldots \quad u'_{m,t} \quad v'_{m,t} \quad \ldots \quad u'_{N,t} \quad v'_{N,t} \big]^T \\ [U_{m,t}]^k = [U_{m,1}]^i [l'_m]_j \mathrm{T}_i^{jk} & , \quad [U'_{m,t}]^k = [U_{m,1}]^i [l'_m]_j \mathrm{T}'^{jk}_i \end{aligned} \quad (6)$$

The above equations are written in tensor notation. Using the image measurements from the first stereo image pair in the sequence, the 3-D motion $M_t$, together with the extrinsic parameters of the stereo rig $E$, the estimated coordinates of the feature points at current time $t$ can be computed by the transfer function $g_t(M_t)$. T and T' are known as the trifocal tensors [9]. They constrain altogether four views in the measurement model at a time instance as follows. 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 $t$.

In (6), $U_{m,t}$ and $U'_{m,t}$ are respectively the normalized homogenous form of $p_{m,t}$ and $p'_{m,t}$ such that $U_{m,t} = \begin{bmatrix} \bar{u}_{m,t} & \bar{v}_{m,t} & \bar{w}_{m,t} \end{bmatrix}^T = \begin{bmatrix} u_{m,t}/f & v_{m,t}/f & 1 \end{bmatrix}^T$ and $U'_{m,t} = \begin{bmatrix} \bar{u}'_{m,t} & \bar{v}'_{m,t} & \bar{w}'_{m,t} \end{bmatrix}^T = \begin{bmatrix} u'_{m,t}/f & v'_{m,t}/f & 1 \end{bmatrix}^T$. $l'_m$ is a line passing through the $m^{\text{th}}$ feature point in the 1st image of the right view. The details for the computation of tensors T, T', and the construction of line $l'_m$ can refer to [9]. From the dynamic system (4) and measurement model (5), the equations for the prediction and smoothing of state estimates in the filter can be derived. Details can be found in [7].

The procedure to handle the changeable set of point features is simple. 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 KLT tracker will be bootstrapped.

## 4. An empirical comparison with existing methods

In the comparison, we investigated the 3-D motion errors of our algorithm against the increase in camera motion. We varied each of the rotation angles from 0.1 to 2.6 degrees per frame and each of the translation parameters from 0.001 to 0.026 meters per frame synchronously. The synthetic structure consisted of

150 random feature points. The focal length of the camera was 6mm with a 2-D zero-mean Gaussian noise of 1 pixel standard deviation. 50 independent tests for each case were carried out. The proposed algorithm, the EKF by Azarbayejani and Pentland [5] and the 2-step EKF by Yu *et. al.* [1] were implemented in Matlab and run on a Pentium IV 2GHz machine to estimate the camera motion. To ensure fair comparison, the number of measurements input to the EKFs was equal.

Fig. 2 shows the results. You can see that the errors of the proposed approach remained at a low level even when the motion was increasing. Actually, the proposed algorithm, the EKF by Azarbayejani and Pentland and Yu's 2-step EKF took 0.130s, 0.150s and 0.061s to process an extra image frame/pair, respectively. Yu's 2-step EKF was the most computation efficient because a tradeoff between speed and accuracy was made in their implementation.

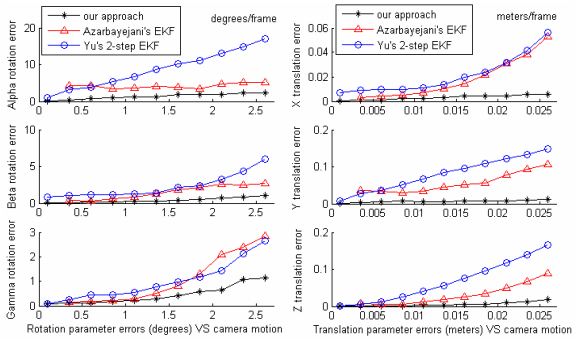


Fig. 2. The relation of the rotation and translation parameter errors against the increase in camera motion.

## 5. Application to super-resolution

To further demonstrate the validity of our motion estimation algorithm, a real stereo image sequence, which is composed of 115 frames, was used to test the proposed approach. The motion of the sequence consisted of rotations and translations on the x-z plane. The extracted 3-D motion was applied to produce super-resolved stereo images.

The algorithm similar to that of Irani and Peleg [13] was used to enhance the resolution. Before the enhancement procedure can be applied, a full point-to-point correspondence between two stereo image pairs $p_{m,t}$, $p'_{m,t}$ and $p_{m,t+b}$, $p'_{m,t+b}$, where $b$ is an integer, for image warping is required. One way to obtain a stereo correspondence map is through the techniques of optic flow or graph cut [16]. Then the mapping of every pixel between any two stereo pairs can be uniquely determined using the two trifocal tensors that relate every four views in the sequence. However, the computation involved in finding dense correspondences with optic flow or graph cut is tedious [16].

Instead, we propose a simple and effective approach to achieve the task. The idea is to first compute the correspondences of the corner features between image pairs $p_{m,t}$, $p'_{m,t}$ and $p_{m,t+b}$, $p'_{m,t+b}$ using the trifocal tensors. A scattered fitting is then performed to reconstruct the estimation for the x- and y- mapping of other pixels between the two pairs of image frames. Bicubic interpolation is used in the fitting. The fitted surface can thus provide a dense correspondence for the two pairs of images. The advantage of adopting surface fitting is that we are using the reliable corner features to guide the interpolation.

The correspondence map takes the role of the global image registration step in the algorithm by Irani and Peleg [13]. In this way, the image warping parameters consist of the feature correspondences predicted by our motion estimation algorithm. Finally, the Irani-Peleg re-projection algorithm [13] is performed to iteratively refine the stereo images until no improvement can be observed.

Figs. 3 and 4 show the results. First of all, we verified whether the recovered 3-D motion was correct. Corner features in the 1st image pair of the stereo sequence were extracted and re-projected to the succeeding frames using the motion parameters recovered. We checked the consistency of the motion of these corner features with respect to the background images. From Fig. 3, you can see that the features, which are indicated by cross markers, stick to the same position relative to the background in these two pairs of images.

Results of super-resolution are shown in Fig. 4. We have implemented the original algorithm by Irani and Peleg [13] that assumes planar image rotation and translation as a comparison. The original Irani-Peleg implementation was unable to correctly enhance the region shown due to the 3-D rotation motion as well as the movement along the z-axis. Significant aliasing could be seen. On the contrary, our method could correctly enhance the image quality with more details revealed from the texts on the book covers.

## 6. Conclusion

We propose a recursive 3-D motion estimation algorithm for a stereo camera set up based on the trifocal tensor. Its high performance in speed and accuracy has been verified in the empirical comparison with other SAM-based methods. The proposed algorithm has been applied to registering 2-D images through the 3-D space for the construction of super-resolved images. Significant improvement on the

image quality is found compared to the traditional methods that are based on 2-D image motion.

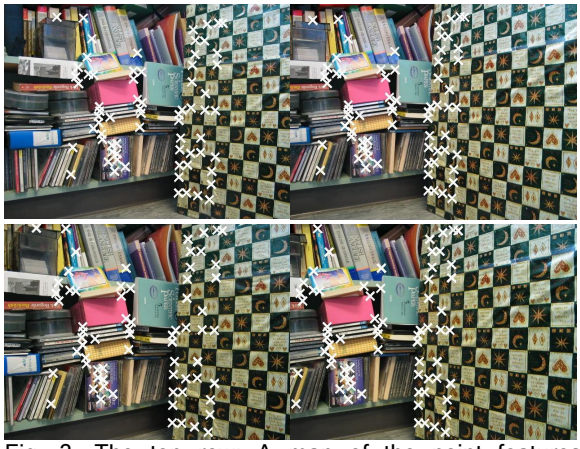

Fig. 3. The top row: A map of the point features extracted from the 1st image pair of the stereo sequence. The bottom row: The re-projection of the point features in the 105th image pair.


Fig. 4. A comparison of the resulting super-resolved images. The enlarged region located near the middle top of the left view is shown. The upper and the lower picture are the results by Irani's algorithm and our approach, respectively.

## 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.

## 8. References

[1] 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.

[2] 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.

[3] 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.

[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] 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.

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

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

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

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

[11] S.Soatto, R.Frezza and P.Perona, "Motion estimation on the essential manifold", presented at ECCV, Stockholm, Sweden, May 1994.

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

[13] M.Irani and S.Peleg, "Improving resolution by image registration", *CVGIP: Graph. Models Image Process.*, vol. 53, pp. 231-239, Mar. 1991.

[14] M.Elad and A.Feuer, "Restoration of single super-resolution image from several blurred, noisy and down-sampled measured images", *IEEE Trans. Image Processing*, vol. 6, no. 12, pp. 1646-1658, Dec. 1997.

[15] S.Farsiu, D.Robinson, M.Elad and P.Milanfar, "Fast and robust multi-frame super-resolution", *IEEE Trans. Image Processing,* vol. 13, no. 10, pp. 1327-1344, Oct. 2004.

[16] W.Zhao and H.Sawhney, "Is super-resolution with optical flow feasible?", in *Proc. ECCV*, vol. 1, pp. 599-613, Copenhagen, May 2002.

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

[18] 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.