SIMULTANEOUS STEREO-MOTION FUSION AND 3-D MOTION TRACKING

Yucel Altunbasak, A. Murat Tekalp and Gozde Bozdagi

Department of Electrical Engineering and Center for Electronic Imaging Systems,
University of Rochester, Rochester, NY 14627, altunbas@ee.rochester.edu

ABSTRACT

We present a new framework for combining maximum likelihood (ML) stereo-motion fusion with adaptive iterated extended Kalman filtering (IEKF) for 3-D motion tracking. The ML stereo-fusion step, with two stereo-pairs, generates observations of 3-D feature matches to be used by the IEKF step. The IEKF step, in turn, computes updated 3-D motion parameter estimates to be used by the ML stereo-motion fusion step. The covariance of the observation noise process is regulated by the value of the ML cost function to address occlusion related problems. The proposed simultaneous approach is compared with performing the 3-D feature correspondence estimation and the Kalman filtering separately using simulated stereo imagery.

1. INTRODUCTION

Several researchers have proposed extended Kalman filtering (EKF) for tracking 3-D motion parameters from long stereo sequences using 2-D or 3-D feature correspondences as observations [1], [2], [3]. These formulations treat the estimation of the 2-D or 3-D feature correspondences from pairs of frames and tracking of the 3-D motion parameters more-or-less separately. Statistical data association techniques [5] or deterministic stereo-motion fusion using dynamic programming [6] have been used in the literature to estimate 3-D feature matches. However feature matching is itself an ill-posed problem, and without some stronger regularization constraints erroneous matches may be found, resulting in the divergence of the EKF. To this effect, in this paper we propose a simultaneous framework which combines the maximum likelihood (ML) estimation of 3-D feature correspondences with extended Kalman filtering for 3-D motion tracking.

The ML correspondence estimation algorithm imposes consistency of the estimated feature matches with the projected 3-D motion parameters obtained from the EKF. However, unlike the previous approaches where the feature matching step is initialized by the predicted estimate from the EKF, the new ML approach imposes this consistency constraint as part of the cost function. At each time instant, several iterations of the ML and the EKF algorithms are interleaved such that the EKF makes use of the feature correspondences computed by the ML step, and the ML step uses the 3-D motion parameters updated by the EKF. That is, we have an iterated EKF (IEKF) which uses improved observations at each iteration as well as an improved linearization. The algorithm advances to the next time sample when the ML cost function can no longer be reduced. In section II, we discuss the stereo imaging geometry and the motion kinematics. In Section III, the ML step of the algorithm is introduced. The formulation of an adaptive EKF, where the variances of the observation noise are adjusted by the value of the ML cost function to address occlusion related problems, is discussed in Section IV. In Section V, we compare the results of the combined algorithm with those obtained by treating the two-steps separately.

This work is supported in part by a National Science Foundation IUCRC grant and a New York State Science and Technology Foundation grant to the Center for Electronic Imaging Systems at the University of Rochester, and a grant by Eastman Kodak Company.

2. IMAGING AND MOTION MODELS

Imaging Model: Fig. 1 shows a 3-D world coordinate system $\mathbf{C}_W$ and two camera coordinate systems $\mathbf{C}_L$ and $\mathbf{C}_R$ that are fixed on the left and right cameras, respectively, with their z-axes pointing along the optical axis of the cameras. Let $f_l = f_r = f$ denote the focal length of both cameras. Consider two other coordinate systems $\mathbf{C}_O$, the object coordinate system, whose origin $\mathbf{P}_O$ coincides with the center of rotation (that is unknown), and $\mathbf{C}_S$, the structure coordinate system, whose origin is located at a known point on the object. It is assumed that $\mathbf{C}_S$ and $\mathbf{C}_O$ are related by a translation $\mathbf{d}$.

Let a point $\mathbf{P}_i(t) = (X_i(t), Y_i(t), Z_i(t))$ in the world coordinate system $\mathbf{C}_W$ be represented as $\mathbf{P}_{i_r}(t) = (X_{i_r}(t), Y_{i_r}(t), Z_{i_r}(t))$ in $\mathbf{C}_R$ and $\mathbf{P}_{i_l}(t) = (X_{i_l}(t), Y_{i_l}(t), Z_{i_l}(t))$ in $\mathbf{C}_L$, respectively. Then,

$$\mathbf{P}_{i_r}(t) = \mathbf{R}_r \mathbf{P}_i(t) + \mathbf{T}_r, \tag{1}$$

$$\mathbf{P}_{i_l}(t) = \mathbf{R}_l \mathbf{P}_i(t) + \mathbf{T}_l, \tag{2}$$

where $\mathbf{R}_r$ and $\mathbf{R}_l$ represent the rotation matrices, and $\mathbf{T}_r$ and $\mathbf{T}_l$ denote the translation vectors indicating the

0-7803-2431-5/95 $4.00

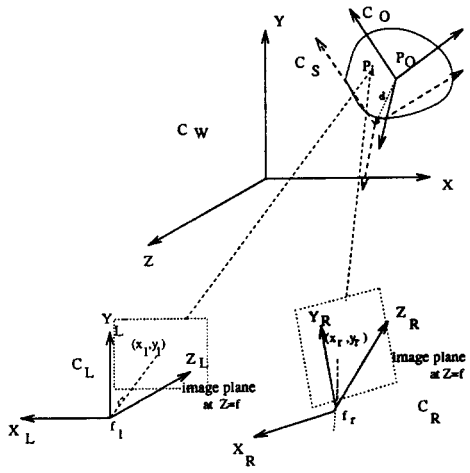

Figure 1: Geometry of stereo imaging

relative positions of $\mathbf{C}_R$ and $\mathbf{C}_L$ with respect to $\mathbf{C}_W$, respectively. Combining equations (1) and (2), we have

$$\begin{aligned}\mathbf{P}_{i_l}(t) &= \mathbf{R}_l\mathbf{R}_r^{-1}\mathbf{P}_{i_r}(t) - \mathbf{R}_l\mathbf{R}_r^{-1}\mathbf{T}_r + \mathbf{T}_l \\ &= \mathbf{M}\mathbf{P}_{i_r}(t) + \mathbf{B}, \end{aligned} \quad (3)$$

where $\mathbf{M}$ and $\mathbf{B}$ are known as camera calibration matrices. Then, the perspective projection of the point $\mathbf{P}_i$ into the left and right image planes can be expressed as

$$\begin{aligned} x_{i_l}(t) = f\frac{X_{i_l}(t)}{Z_{i_l}(t)} \quad y_{i_l}(t) = f\frac{Y_{i_l}(t)}{Z_{i_l}(t)}, \\ x_{i_r}(t) = f\frac{X_{i_r}(t)}{Z_{i_r}(t)} \quad y_{i_r}(t) = f\frac{Y_{i_r}(t)}{Z_{i_r}(t)}, \end{aligned} \quad (4)$$

respectively. Substituting (4) into (3), we have

$$\frac{Z_{i_l}(t)}{f}\begin{bmatrix} x_{i_l}(t) \\ y_{i_l}(t) \\ f \end{bmatrix} = \frac{Z_{i_r}(t)}{f}\mathbf{M}\begin{bmatrix} x_{i_r}(t) \\ y_{i_r}(t) \\ f \end{bmatrix} + \mathbf{B}. \quad (5)$$

Given the horizontal disparity, $d_{x_i}(t) = x_{i_r}(t) - x_{i_l}(t)$, and the vertical disparity, $d_{y_i}(t) = y_{i_r}(t) - y_{i_l}(t)$, we can find the 3-D world coordinates of the corresponding object point using (1)-(5).

Motion model: We use the motion model of Young and Chellappa [4] in our formulations. The translational component of motion is represented by a constant acceleration model given by

$$\begin{aligned} \mathbf{P}_O(t_{k+1}) = \mathbf{P}_O(t_k) + (t_{k+1} - t_k)\mathbf{v}(t_k) + \frac{1}{2}(t_{k+1} - t_k)^2\mathbf{a}(t_k), \\ \mathbf{v}(t_{k+1}) = \mathbf{v}(t_k) + (t_{k+1} - t_k)\mathbf{a}(t_k), \\ \mathbf{a}(t_{k+1}) = \mathbf{a}(t_k), \end{aligned} \quad (6)$$

where $\mathbf{v}(t_k) = (v_x(t_k), v_y(t_k), v_z(t_k))$ and $\mathbf{a}(t_k) = (a_x(t_k), a_y(t_k), a_z(t_k))$ denote the translational velocity and acceleration vectors, respectively. The rotational component of motion will be represented by a quaternion $\mathbf{q}(t_k) = (q_1(t_k), q_2(t_k), q_3(t_k), q_4(t_k))$ under the assumption of a constant precession model. It has been shown that the temporal dynamics of the unit quaternion can be expressed in closed form as a function of ω and $\mathbf{p}$, where $\omega = (\omega_x, \omega_y, \omega_z)$ and $\mathbf{p} = (p_x, p_y, p_z)$ denote the angular velocity and precession vectors, respectively [4].

3. ML STEREO-MOTION FUSION

In this section, we present the ML feature correspondence estimation step of the simultaneous algorithm. Let

$$\begin{aligned} p(\mathbf{I}_L(t_{k+1}), \mathbf{I}_R(t_{k+1}), \mathbf{I}_L(t_k) | \mathbf{u}_{l_x}(t_k), \mathbf{u}_{l_y}(t_k), \mathbf{u}_{r_x}(t_k), \\ \mathbf{u}_{r_y}(t_k), \mathbf{d}_x(t_k), \mathbf{d}_y(t_k), \mathbf{I}_R(t_k)) \end{aligned} \quad (7)$$

denote the conditional probability distribution (pdf) of $\mathbf{I}_L(t_{k+1}), \mathbf{I}_R(t_{k+1})$, the left and right images at time t_{k+1}, and $\mathbf{I}_L(t_k)$, left image at time t_k, given $\mathbf{u}_{l_x}(t_k)$, $\mathbf{u}_{l_y}(t_k), \mathbf{u}_{r_x}(t_k), \mathbf{u}_{r_y}(t_k)$, vectors formed by lexicographic ordering of the components of the vertical and horizontal motion vectors between the left image pairs, and the right image pairs, respectively; $\mathbf{d}_x(t_k)$ and $\mathbf{d}_y(t_k)$, lexicographic ordering of horizontal and vertical disparity vectors at time t_k at all feature points; and $\mathbf{I}_R(t_k)$ the right image at time t_k. The ML estimates of the 2-D motion and disparity vectors are those that maximize the conditional pdf (7).

The conditional probability distribution (7) provides a measure of how well the present motion and disparity estimates conform with the observed frames $\mathbf{I}_L(t_{k+1})$, $\mathbf{I}_R(t_{k+1}), \mathbf{I}_L(t_k)$ given the frame $\mathbf{I}_R(t_k)$. In the following, it is modeled by a Gibbsian distribution, given by

$$\begin{aligned} p(\mathbf{I}_L(t_{k+1}), \mathbf{I}_R(t_{k+1}), \mathbf{I}_L(t_k) | \mathbf{u}_{l_x}(t_k), \mathbf{u}_{l_y}(t_k), \\ \mathbf{u}_{r_x}(t_k), \mathbf{u}_{r_y}(t_k), \mathbf{d}_x(t_k), \mathbf{d}_y(t_k), \mathbf{I}_R(t_k)) = \\ \frac{1}{Z_1} exp(-U(\mathbf{I}_L(t_{k+1}), \mathbf{I}_R(t_{k+1}), \mathbf{I}_L(t_k) | \mathbf{u}_{l_x}(t_k), \\ \mathbf{u}_{l_y}(t_k), \mathbf{u}_{r_x}(t_k), \mathbf{u}_{r_y}(t_k), \mathbf{d}_x(t_k), \mathbf{d}_y(t_k), \mathbf{I}_R(t_k)), \end{aligned}$$

where Z is a constant, and $U(.)$ is given by

$$\begin{aligned} U(\mathbf{I}_L(t_{k+1}), \mathbf{I}_R(t_{k+1}), \mathbf{I}_L(t_k) | \mathbf{u}_{l_x}(t_k), \mathbf{u}_{l_y}(t_k), \\ \mathbf{u}_{r_x}(t_k), \mathbf{u}_{r_y}(t_k), \mathbf{d}_x(t_k), \mathbf{d}_y(t_k), \mathbf{I}_R(t_k)) = \\ \sum_{i=1}^{N}[\alpha_1\epsilon_1(i) + \alpha_2\epsilon_2(i) + \alpha_3\epsilon_3(i) + \alpha_4\epsilon_4(i)], \end{aligned} \quad (8)$$

N is the number of feature points, and

$$\epsilon_1(i) = \sum_{(m,n)\in\mathcal{N}_i} |I_L(m + d_x(m,n), n + d_y(m,n); t_k) - I_R(m,n;t_k)|^2,$$

$$\epsilon_2(i) = \sum_{(m,n)\in\mathcal{N}_i} |I_L(m'_L, n'_L; t_{k+1}) - I_R(m,n;t_k)|^2,$$

$$\epsilon_3(i) = \sum_{(m,n)\in\mathcal{N}_i} |I_R(m+u_x(m,n), n+u_y(m,n); t_{k+1}) - I_R(m,n;t_k)|^2,$$

$$\epsilon_4(i) = \sum_{(m,n)\in\mathcal{N}_i} \left[||\mathbf{u}_r(m,n;t_k) - \tilde{\mathbf{u}}_r(m,n;t_k)||^2 + ||\mathbf{u}_l(m,n;t_k) - \tilde{\mathbf{u}}_l(m,n;t_k)||^2 \right]$$

Here, $\mathcal{N}_i$ denotes a neighborhood of the feature i, (m'_L, n'_L) refers to the coordinates in the left image at time t_{k+1} that corresponds to (m,n) in the right image at time t_k, and $\tilde{\mathbf{u}}_r(m,n;t_k)$ and $\tilde{\mathbf{u}}_l(m,n;t_k)$ are the projected motion vectors at pixel (m,n) at time t_k obtained as described in the following algorithm. Observe that the last term in the potential function (8) enforces consistency of the correspondence estimates with the projected 3-D motion [7].

Algorithm: The maximization of (7) is pursued in the following manner:

1. Select N feature points in the initial frame t_0. Initialize the disparity vectors $\mathbf{d}_x(t_0)$ and $\mathbf{d}_y(t_0)$, and the motion vectors $\mathbf{u}_{l_x}(t_0)$, $\mathbf{u}_{l_y}(t_0)$, $\mathbf{u}_{r_x}(t_0)$, $\mathbf{u}_{r_y}(t_0)$ using a block correlation method between the respective frames.

   Find the 3-D coordinates of each feature point $\mathbf{P}_i(t_0)$ using (1)-(5). Set $k=0$.

2. Find the 3-D coordinates of each feature point $\mathbf{P}_i(t_{k+1})$ given $\mathbf{u}_{l_x}(t_k)$, $\mathbf{u}_{l_y}(t_k)$, $\mathbf{u}_{r_x}(t_k)$, $\mathbf{u}_{r_y}(t_k)$ using (1)-(5).

3. Given the 3-D point matches $\mathbf{P}_i(t_k)$ and $\mathbf{P}_i(t_{k+1})$ as observables, estimate the 3-D motion parameters $\mathbf{q}(t_k)$, $\mathbf{w}(t_k)$ and $\mathbf{v}(t_k)$ using EKF (see Section 4).

4. Find the projected motion vectors $\tilde{\mathbf{u}}_{r_x}(t_k)$, $\tilde{\mathbf{u}}_{r_y}(t_k)$, $\tilde{\mathbf{u}}_{l_x}(t_k)$ and $\tilde{\mathbf{u}}_{l_y}(t_k)$ as follows: Apply the 3-D motion parameters $\mathbf{R}$ and $\mathbf{T}$ to $\mathbf{P}_i(t_k)$ to find $\tilde{\mathbf{P}}_i(t_{k+1})$. Project $\tilde{\mathbf{P}}_i(t_{k+1})$ into the left and right image planes, and take the difference between the respective image points at time t_k.

5. Update the 2-D motion and disparity vectors by a gradient descent algorithm to minimize

$$\text{Cost}(t_k) = \sum_{i=1}^{N} \text{Cost}_i(t_k) \tag{9}$$

where

$$\text{Cost}_i(t_k) = U(\mathbf{I}_L(t_{k+1}), \mathbf{I}_R(t_{k+1}), \mathbf{I}_L(t_k) | \mathbf{u}_{l_x}(t_k), \mathbf{u}_{l_y}(t_k), \mathbf{u}_{r_x}(t_k), \mathbf{u}_{r_y}(t_k), \mathbf{d}_x(t_k), \mathbf{d}_y(t_k)). \tag{10}$$

is evaluated on a local window centered about the feature point i.

We iterate through steps 2-5 at time t_k until convergence.

6. Increment k by 1 and set $\mathbf{u}_{l_x}(t_k)$, $\mathbf{u}_{l_y}(t_k)$, $\mathbf{u}_{r_x}(t_k)$, $\mathbf{u}_{r_y}(t_k)$ to the predicted values from the EKF and go to Step 2.

4. EXTENDED ADAPTIVE KALMAN FILTERING

Due to the nonlinearities in the state and the measurement models caused by the relationship between the quaternion representation and the rotation parameters, we use an extended Kalman filter (EKF) method to estimate the 3-D motion parameters as proposed by Young and Chellappa [4].

The observations of the EKF are the 3-D feature correspondences $\mathbf{P}_i(t_k)$ and $\mathbf{P}_i(t_{k+1})$ found by the ML step in Section 3. Given the imaging and motion model described in Section 2, the state transition and the measurement equations for the Kalman filter are

State equation: $\mathbf{s}(t_{k+1}^-) = f(\mathbf{s}(t_k^+), \omega(t_k^+), \mathbf{p}(t_k^+))$, where

$$\mathbf{s}(t_k) = \left[\begin{array}{ccccccc} \mathbf{P}_O(t_k) & \mathbf{v}(t_k) & \mathbf{a}(t_k) & \mathbf{d}(t_k) & \mathbf{q}(t_k) & \omega(t_k) & \mathbf{p}(t_k) \end{array} \right]^T$$

and

Measurement equation:

$$\mathbf{P}_i(t_k) = \mathbf{P}_O(t_k) + \mathbf{R}\left[\,\mathbf{q}(t_k)\,\right](\mathbf{P}_{s_i}(t_k) - \mathbf{d}) + \mathbf{n}(t_k)$$

where $\mathbf{R}$ is the rotation matrix defined with respect to the center of rotation, $\mathbf{P}_s(t_k)$ is the 3-D coordinates of the feature point with respect to C_S, and $n(t_k)$ is the measurement noise which is assumed to be zero mean with covariance matrix $\mathbf{N}(t_k)$.

The covariance matrix $\mathbf{N}(t_k)$ can be adaptively modified to handle problems with occlusion. We adjust the diagonal elements of the covariance matrix by the ML cost function given by (10) as

$$\mathbf{N}(t_k) = \begin{bmatrix} \sigma_1^2(t_k) & 0 & \cdots & 0 \\ 0 & \sigma_2^2(t) & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & & \sigma_N^2(t_k) \end{bmatrix} \tag{11}$$

where

$$\sigma_i^2(t_k) = \sigma_0^2(t_k) + \rho(t_k) Cost_i(t_k), \tag{12}$$

$\rho(t_k)$ is a scalar weight, and $\sigma_0^2(t)$ is the nominal noise variance. When a feature point i is occluded, then a correspondence can not be found for this point which causes $Cost_i(t_k)$, hence $\sigma_i^2(t_k)$ to be large. This adaptivity decreases the effect of occlusion in the tracking performance of the EKF.

in cm.	Feature 1	Feature 2
True depth	100.00	102.47
Block corr.	98.70	102.08
Dyn. prog.	99.32	102.31
ML	99.46	102.41

Table 1: Feature estimation performance

5. RESULTS

In this section, we compare i) the feature matching performance of the ML step with a deterministic algorithm [6], ii) the tracking performance of the adaptive versus non-adaptive ML-EKF algorithm, and iii) the tracking performance of the adaptive EKF with feature correspondences obtained by the ML step versus by the deterministic algorithm. Results are provided on a simulated stereo image sequence consisting of 40 frames. The consecutive frames are generated by rotating a texture mapped rectangular block by the 3-D rotation parameters, $\omega_x = -1$, $\omega_x = -1$, $\omega_x = -3$, and the 3-D translation parameters $T_x = -0.5$, $T_y = 0.5$ and $T_z = 0.5$ where the rotations are in degrees and the translations are in centimeters. Occlusion is simulated by moving another rectangular block over the image. We select 8 feature points on the right image at time t_0.

To evaluate the feature matching performance, we have computed the depth of two feature points using three different algorithms. The results shown in Table 1 indicate that the ML algorithm outperforms both of the deterministic techniques. Note that this experiment indeed evaluates only disparity estimates.

In order to provide a comparison of the tracking performance of the ML-EKF algorithm, we have plotted the magnitude of error in the rotational and translational velocity vectors over 40 frames in Figs. 2 and 3, respectively. In the figures, "ML-2 view" refers to estimating the 3-D motion parameters directly from the 2-D point matches by a weighted least squares approach on a frame-by-frame basis. In "non-adaptive EKF" the observation noise covariance matrix is fixed for all t_k. "Dyn-EKF" refers to running the EKF with the feature correspondences obtained from the deterministic algorithm. The results demonstrate the superior performance of the ML-EKF algorithm.

6. CONCLUSION

We presented a new formulation where feature matching and 3-D motion tracking are performed in a single iterative framework. The algorithm alternates between an ML step and IEKF step at each iteration until the ML cost function can no longer be reduced. Results prove superior to performing each step separately.

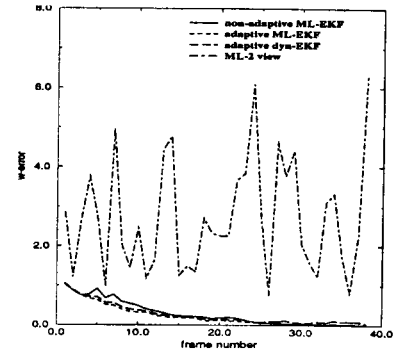

Figure 2: Error in rotation parameters

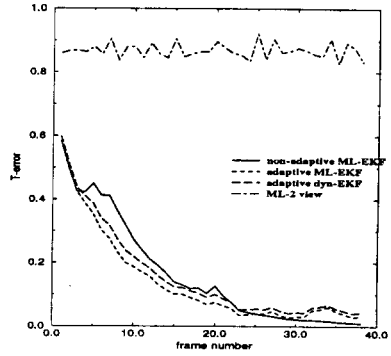

Figure 3: Error in translation parameters

7. REFERENCES

[1] S. Lee and Y. Kay, "A Kalman Filter Approach for Accurate 3-D Motion Estimation from a Sequence of Stereo Images," *CVGIP: Im. Understanding.* vol. 54, No. 2, pp. 244-258, 1991.

[2] Y. Yao and R. Chellappa, "Dynamic feature point tracking in an image sequence," *Int. Conf. Pattern Rec.*, pp. 654–657, Oct. 1994.

[3] Z. Zhang and O. D. Faugeras, "Three-Dimensional Motion Computation and Object Segmentation in a Long Sequence of Stereo Frames," *Int. Journal of Comp. Vision*, vol. 7, pp.211-241, 1992.

[4] G. J. Young and R. Chellappa, "3-D Motion Estimation Using a Sequence of Noisy Stereo Images : Models, Estimation, and Uniqueness Results", *IEEE Trans. Pat. Anal. Mach. Intell.*, vol. 12, No. 8, 1990.

[5] I. J. Cox, "A review of statistical data association techniques for motion correspondence," *Int. Journal of Comp. Vision*, vol. 10, no. 1, pp. 53–66, 1993.

[6] J. Liu and R. Skerjanc,"Stereo and motion correspondence in a sequence of stereo images," *Signal Proc.: Image Comm.*, vol. 5, no. 4, pp.305–318, Oct. 1993.

[7] Y. Altunbasak, A. M. Tekalp and G. Bozdagi, "Simultaneous Motion-Disparity Estimation and Segmentation From Stereo," *IEEE Int. Conf. on Image Proc.*, (Austin, TX), November 10-12, 1994