Comparison of Robust Estimation and Kalman Filtering Applied to Fingertip Tracking in Human-Machine Interfaces

Sylvia M. Dominguez
Electrical Engineering Dept.
University of California
Los Angeles, CA 90095

Trish Keaton
Information Sciences Lab.
HRL Laboratories LLC
Malibu, CA 90265

Ali H. Sayed*
Electrical Engineering Dept.
University of California
Los Angeles, CA 90095

Abstract

This paper studies the application of robust state-space estimation with uncertain models to tracking problems in human-machine interfaces. The need for robust methods arises from the desire to control the influence of uncertain environment conditions on system performance, such as the effect of abrupt variations in object speed and motion characteristics. This paper produces models for motion uncertainties associated with a human hand, and applies them to a robust state-space estimation algorithm used to track a user's pointing fingertip. Then a comparison is performed between the results from the robust tracker against a Kalman filter.

1 Introduction

One critical factor in human-machine interface applications is the ability of the machine to quickly and efficiently identify and interpret the hand gestures of its user. This capability can be useful in many circumstances. For example, while using a wearable computer system, the user's fingertip could be used to point to and encircle objects of interest in a scene. In this way, a machine that is able to track the movements of the user's fingertip could convey to the user information about the identified objects.

There are several computer vision algorithms that have been developed for such purposes in the literature [1],[7]. These algorithms extract color segmentations, 3D stereo segmentations, and shape information from the machine's camera view in order to identify the user's hand and fingertip position. The algorithms, however, are complex and computationally intensive, and tend to slow down the response of the machine to a great extent. In order to perform real-time acquisition and tracking, state-space estimation techniques can be used to enable the designer to reduce the search space from the full camera view to a smaller search window in a dynamic fashion. This window is centered around a prediction for the future position of the object being tracked, and such predictions can be obtained from state-space estimation methods like, e.g., Kalman filters (refer to Figures 1 and 2).

However, since the trajectory created by the user's hand is subject to several sources of uncertainties, it becomes useful to investigate the use of robust estimation methods in order to limit the degradation in performance of otherwise optimal systems. For a wearable computer system, for example, these uncertainties arise from the camera moving along with the user's head motion, the background and object moving independently of each other, the user standing still or randomly walking, and the user's pointing finger abruptly changing directions at variable speeds. All these factors give rise to uncertainties that influence the design of reliable trackers. This paper attempts to model such sources of uncertainties and compares the performance of a tracker that is based on a Kalman filter to that of a tracker that is based on the robust algorithm of Sayed [6]. The latter shows improved performance.

2 State-space modeling for fingertip tracking

Figure 1 illustrates a wearable computer system, *Snap&Tell*, that is currently under development at HRL laboratories in Malibu,CA. It aims at providing gesture-based interfaces between users and machines. This system enables a user to specify, segment, and recognize objects of interest, such as landmarks, by simply pointing at and encircling them with the user's fingertip (for details see [2],[3]).

In order to enhance the gesture-based interface, a

*The work of A. H. Sayed was partially supported by NSF award ECS-9820765.

0-7803-7147-X/01/$10.00©2001 IEEE

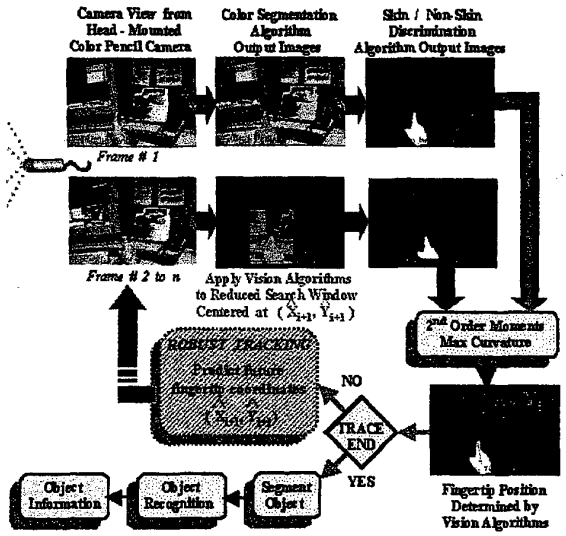

Figure 1: *Block diagram of gesture-based interface for a wearable computer system.*

state-space tracker is inserted into the system in order to predict the future user's fingertip position, which will point to the user's focus of attention during the next frame. These predicted coordinates are then used as the center of a smaller image search window during the next video frame, thus speeding up the response time of the system and making it memory and computationally efficient.

The tracker relies on a state-space model that describes the fingertip motion. Thus let T denote the frame capture rate for the wearable computer system (measured in seconds/frame). Let also $\{\alpha_{x,i}, \alpha_{y,i}\}$ denote the fingertip's accelerations along the x and y directions (measured in pixels per second^2), and let $\{v_{x,i}, v_{y,i}\}$ denote the speeds along these same directions during the i^{th} frame (measured in pixels/second). Then one could approximate the present fingertip position in the i^{th} frame $\{x_i, y_i\}$ in terms of the previous frame fingertip pixel coordinates $\{x_{i-1}, y_{i-1}\}$ and the pixel-shifts per frame $\{v_{i-1}T, \alpha_{i-1}\frac{T^2}{2}\}$ such as

$$v_{x,i} \approx v_{x,i-1} + \alpha_{x,i-1}T \quad (1)$$

$$v_{y,i} \approx v_{y,i-1} + \alpha_{y,i-1}T \quad (2)$$

$$x_i \approx x_{i-1} + v_{x,i-1}T + \alpha_{x,i-1}\frac{T^2}{2} \quad (3)$$

$$y_i \approx y_{i-1} + v_{y,i-1}T + \alpha_{y,i-1}\frac{T^2}{2} \quad (4)$$

In general, the acceleration and speed variables are unknown and more advanced trackers would need to estimate them as well. In order to simplify the presentation in this article a simplified model is adopted. Thus assume zero accelerations in the x and y directions and, consequently, constant speeds along these directions. These assumptions are fairly reasonable in situations when the user is standing still and pointing at an object.

Under the constant speed assumption, let $\{\Delta x_i, \Delta y_i\}$ denote the pixel displacements in the x and y directions during the i^{th} frame, which will account for the instantaneous change of trajectory of the pointing finger. Then it holds that

$$x_{i+1} = x_i + \Delta x_i, \qquad y_{i+1} = y_i + \Delta y_i \quad (5)$$

and these equations motivate the following state-space model. Introduce the state and measurement vectors

$$s_i \triangleq \begin{bmatrix} x_i \\ y_i \\ \Delta x_i \\ \Delta y_i \end{bmatrix}, \quad z_i \triangleq \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (6)$$

then

$$s_{i+1} = Fs_i + Gu_i, \qquad z_i = Hs_i + v_i \quad (7)$$

with model parameters

$$F = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \quad (8)$$

and where u_i and v_i denote uncorrelated zero mean white process and measurement noises that satisfy

$$E\left(\begin{bmatrix} s_0 \\ u_i \\ v_i \end{bmatrix}\begin{bmatrix} s_0 \\ u_j \\ v_j \end{bmatrix}^T\right) = \begin{bmatrix} \Pi_0 & 0 & 0 \\ 0 & Q\delta_{ij} & 0 \\ 0 & 0 & R\delta_{ij} \end{bmatrix} \quad (9)$$

The covariance matrices are chosen as

$$Q = \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad R = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} \quad (10)$$

These choices are determined experimentally as follows. Assuming initially large uncertainties in the

x and y locations, say of the order of 3 pixels, and smaller uncertainties in the displacements, say of the order of 1 pixel. Then, these values are checked for optimality by testing the whiteness of the resulting innovations process of a Kalman filter following the method of Mehra [4]. The chosen values for R and Q meet Mehra's 95% confidence whiteness test.

It should be further mentioned that we are using a full image frame size of 260 × 180, which has an x-pixel mean of 130 and a y-pixel mean of 90. Therefore, these mean values are subtracted from the actual coordinate positions before forming the above state-space model. In other words, the variables $\{x_i, y_i\}$ in the above model are zero-mean variables that have been centered by removing their means. This is needed prior to the application of state-space estimation algorithms, such as a Kalman filter.

In addition, the measurement vector z_i consists of the centered pixel coordinates that are provided by the vision algorithm locating the fingertip position. These coordinates can therefore be regarded as noisy measurements of the actual pixel coordinates $\{x_i, y_i\}$. By using the assumed state-space model (6)–(10), one can then proceed to employ a variety of estimation techniques to 'clean' z_i from measurement noise and to predict future movements of $\{x_i, y_i\}$.

3 Kalman fingertip tracker

Introduce the following predicted and filtered estimates of the state vector:

$$\hat{s}_i \triangleq \text{l.l.m.s. estimate of } s_i \text{ given } \{z_0, z_1, \ldots, z_{i-1}\}$$
$$\hat{s}_{i|i} \triangleq \text{l.l.m.s. estimate of } s_i \text{ given } \{z_0, z_1, \ldots, z_{i-1}, z_i\}$$

and the corresponding error variances,

$$P_i \triangleq E(s_i - \hat{s}_i)(s_i - \hat{s}_i)^T$$
$$P_{i|i} \triangleq E(s_i - \hat{s}_{i|i})(s_i - \hat{s}_{i|i})^T.$$

Then the $\{\hat{s}_i, \hat{s}_{i|i}\}$ can be constructed recursively as follows (see, e.g., [5]):

$$\hat{s}_{i+1} = F\hat{s}_{i|i}, \quad i \geq 0 \tag{11}$$
$$e_{i+1} = z_{i+1} - H\hat{s}_{i+1} \tag{12}$$
$$\hat{s}_{i+1|i+1} = \hat{s}_{i+1} + P_{i+1|i+1}H^T R^{-1} e_{i+1} \tag{13}$$

where

$$P_{i+1} = FP_{i|i}F^T + GQG^T \tag{14}$$
$$R_{e,i+1} = R + HP_{i+1}H^T \tag{15}$$
$$P_{i+1|i+1} = P_{i+1} - P_{i+1}H^T R_{e,i+1}^{-1} HP_{i+1} \tag{16}$$

and with initial conditions

$$\hat{s}_{0|0} = P_{0|0}^{-1} H^T R^{-1} y_0 \tag{17}$$
$$P_{0|0} = (\Pi_0^{-1} + H^T R^{-1} H)^{-1}. \tag{18}$$

Equations (11)–(16) are known collectively as the time- and measurement-update form of the Kalman filter.

Figure 2 shows preliminary results on the tracking of the fingertip location by using the aforementioned state-space model and the Kalman filtering equations. The figure illustrates how the Kalman filter helps reduce the search area and speed up the recognition algorithm. In this particular simulation, the response time of the overall system was reduced by 68% when compared with a system that uses a full camera view to track the user's fingertip. Note how the reduced search window centered around the previously predicted fingertip position, almost overlaps the actual present finger position.

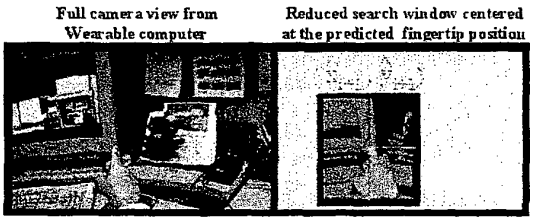

Figure 2: *Successfully tracked fingertip using a Kalman filter.*

It should be noted that the size of the *reduced search window* was chosen to be at least twice the size of the maximum estimation errors in the x and y directions, of a Kalman tracker applied to a training sequence representative of a typical pointing finger trajectory ($\Delta W_x \geq 2\tilde{x}_{max}, \Delta W_y \geq 2\tilde{y}_{max}$). Therefore, the more accurate the tracker is, the smaller the search window needed, and the faster the overall system response time will be. Thus, in order to improve the tracking abilities we now turn to integrate models for the sources of uncertainties, such as abrupt changes in speed or finger trajectory, by using a robust state-space algorithm.

4 Robust fingertip tracker with uncertainty models

It is well-known that a central premise in the Kalman filtering formulation is that it assumes that the underlying model parameters $\{F, G, H, R, Q\}$ are accurate. When this assumption is violated, the performance of the filter can deteriorate appreciably and

one is therefore motivated to consider robust variants; robust in the sense that they attempt to limit, in certain ways, the effect of model uncertainties on the overall filter performance.

For the wearable computer system, there are several sources of uncertainties that may interfere with the accuracy of the assumed state-space model. The uncertainties can be due to the camera moving along with the user's head motion, to changes in lighting conditions, to the background and object moving independently from each other, to the user standing still or randomly walking, or to the user's pointing finger abruptly changing directions at variable speeds and accelerations. All these factors changing constantly in time create different conditions of uncertainties.

One way to model uncertainties is to treat the given parameters $\{F, G\}$ as nominal values and to assume that the actual values lie within a certain set around them. Thus consider an uncertain model of the form

$$\begin{aligned} s_{i+1} &= (F + \delta F_i)s_i + (G + \delta G_i)u_i \quad (19) \\ z_i &= Hs_i + v_i \quad (20) \end{aligned}$$

where the perturbations in $\{F, G\}$ are modeled as

$$\begin{bmatrix} \delta F_i & \delta G_i \end{bmatrix} = M\Delta_i \begin{bmatrix} E_f & E_g \end{bmatrix} \quad (21)$$

for some matrices $\{M, E_f, E_g\}$ and for an arbitrary contraction Δ_i, $\|\Delta_i\| \leq 1$. For generality, one could allow the quantities $\{M, E_f, E_g\}$ to vary with time as well. The model (21) allows the designer to restrict the sources of uncertainties to a certain range space (defined by the matrix M), and to assign different levels of distortion by selecting the entries of $\{E_f, E_g\}$ appropriately — see, e.g., [6]. For example, initial investigations have suggested that possible choices for $\{M, E_f, E_g\}$ in the context of fingertip tracking with constant speed (user standing still) are

$$E_f = \begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix}, \quad E_g = \begin{bmatrix} 1 & 1 & 0 & 0 \end{bmatrix} \quad (22)$$

$$M = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}^T \quad (23)$$

These choices account for possible violations of the constant speed assumption. The authors are currently investigating more elaborate modeling for the uncertainties in surveillance and tracking applications. A listing of a time- and measurement-update form of a robust filter, developed by Sayed [6], that applies to the uncertain model (19)–(21) is shown in Table 1. This robust filter attempts to minimize the estimation error at the worst possible case created by the bounded uncertainties δF_i and δG_i. The estimates $\{\hat{s}_{i+1}, \hat{s}_{i+1|i+1}\}$ and thus the predicted fingertip coordinates for the next frame, can be obtained by recursively iterating between steps 2 and 3. For the case when $\hat{\lambda}_i = 0$, steps 2 and 3 get reduced to the standard time and measurement update Kalman equations.

Assumed uncertain model: Eqs. (19)–(21). Also, $\Pi_0 > 0$, $R > 0$, $Q > 0$ are given weighting matrices.

Initial conditions: Set $\hat{s}_{0|0} = P_{0|0}H^TR^{-1}z_0$ and $P_{0|0} = \left(\Pi_0^{-1} + H^TR^{-1}H\right)^{-1}$.

Step 1. If $HM = 0$, then set $\hat{\lambda}_i = 0$ (non robust filter). Otherwise, select α (typically $0 \leq \alpha \leq 1$) and set

$$\hat{\lambda}_i = (1 + \alpha) \cdot \|M^TH^TR^{-1}HM\|.$$

Step 2. Replace $\{Q, R, P_{i|i}, G, F\}$ by:

$$\begin{aligned} \widehat{Q}_i^{-1} &= Q^{-1} + \hat{\lambda}_i E_g^T \left[I + \hat{\lambda}_i E_f P_{i|i} E_f^T\right]^{-1} E_g \\ \widehat{R}_{i+1} &= R - \hat{\lambda}_i^{-1} HMM^TH^T \\ \widehat{P}_{i|i} &= \left(P_{i|i}^{-1} + \hat{\lambda}_i E_f^T E_f\right)^{-1} \\ &= P_{i|i} - P_{i|i}E_f^T(\hat{\lambda}_i^{-1}I + E_f P_{i|i} E_f^T)^{-1} E_f P_{i|i} \\ \widehat{G}_i &= G - \hat{\lambda}_i F \widehat{P}_{i|i} E_f^T E_g \\ \widehat{F}_i &= (F - \hat{\lambda}_i \widehat{G}_i \widehat{Q}_i E_g^T E_f)(I - \hat{\lambda}_i \widehat{P}_{i|i} E_f^T E_f) \end{aligned}$$

If $\hat{\lambda}_i = 0$, then simply set $\widehat{Q}_i = Q$, $\widehat{R}_{i+1} = R$, $\widehat{P}_{i|i} = P_{i|i}$, $\widehat{G}_i = G$, and $\widehat{F}_i = F$.

Step 3. Update $\{\hat{s}_{i|i}, P_{i|i}\}$ as follows:

$$\begin{aligned} \hat{s}_{i+1} &= \widehat{F}_i \hat{s}_{i|i} \\ e_{i+1} &= z_{i+1} - H\hat{s}_{i+1} \\ P_{i+1} &= F\widehat{P}_{i|i}F^T + \widehat{G}_i\widehat{Q}_i\widehat{G}_i^T \\ R_{e,i+1} &= \widehat{R}_{i+1} + HP_{i+1}H^T \\ P_{i+1|i+1} &= P_{i+1} - P_{i+1}H^TR_{e,i+1}^{-1}HP_{i+1} \\ \hat{s}_{i+1|i+1} &= \hat{s}_{i+1} + P_{i+1|i+1}H^T\widehat{R}_{i+1}^{-1}e_{i+1} \end{aligned}$$

Table 1. *Listing of a robust filtering algorithm in time- and measurement-update form.*

5 Experimental results

We applied this robust algorithm to the fingertip trajectory previously tracked by the plain Kalman filter, using the state model (6)–(10), the perturbation model (22)(23), and the particular choice of α=0.5. The magnitude of the MSE results for both algorithms are shown and compared in Figure 3 for the estimation error of the x and y pixel coordinates, and the estimation error in the Δx and Δy displacements. As it can be seen, the magnitude of the average MSE for

the x, Δx, and Δy estimates are smaller for the robust algorithm when compared with the traditional Kalman filter. However, the average MSE estimate for the y coordinate indicates that the perturbations on (22)(23) did not model properly the y coordinate uncertainties. This prompts us to adjust the perturbation model (23) to increase the level of uncertainty on y, such as

$$M = \begin{bmatrix} 0.1 & 0.2 & 0.1 & 0.1 \end{bmatrix}^T \qquad (24)$$

Figure 4 shows the results obtained with the new perturbation model (22)(24). In this case, the robust algorithm shows smaller magnitudes of the average MSE for all the state variables estimated (x, y, Δx, and Δy), obtaining an averaged improvement of 10% over the overall performance of the traditional Kalman filter.

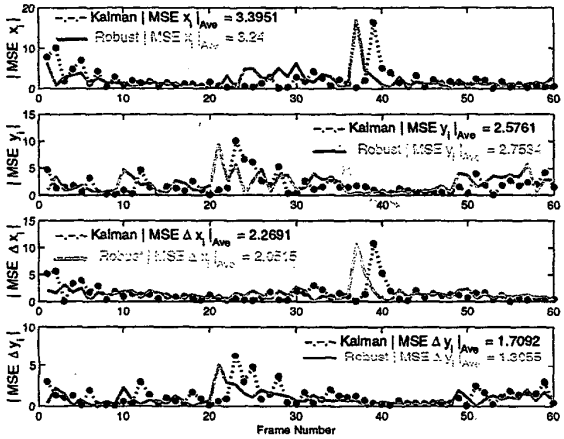

Figure 3: *Comparison of the fingertip estimation errors between the Kalman filter and the robust tracker using the perturbation model (22)(23), with α=0.5.*

6 Conclusion

These performance results are encouraging and merit future exploration. For example, one extension is to investigate on-line adaptive learning methods to develop proper models for uncertainties associated with the user's head motion, walking, and changes in lighting conditions. Another extension is to investigate more elaborate state space variables that allow to model additional information, such as accelerations, depth information, hand size, and skin tone.

References

[1] T. Brown and R.C. Thomas, "Finger tracking for the digital desk", *Proc. Australasian User Interface Conference (AUIC)*, vol. 1, pp. 11-16, Canberra,Australia, 2000.

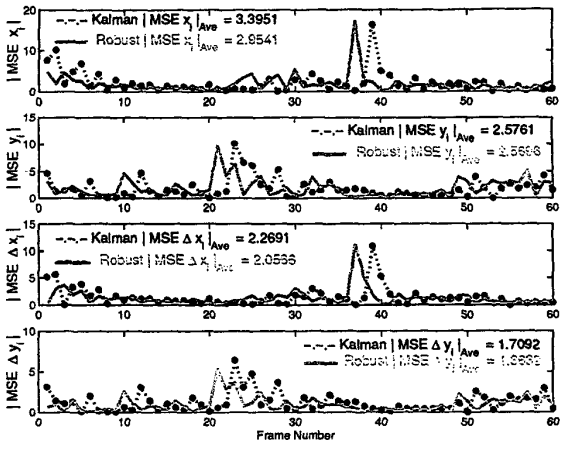

Figure 4: *Comparison of the fingertip estimation errors between the Kalman filter and the robust tracker using the perturbation model (22)(24), with α=0.5.*

[2] S. M. Dominguez, T. A. Keaton, A. H. Sayed, "Robust Finger Tracking for Wearable Computer Interfacing", *Proc. Perceptive User Interfaces (PUI)*, Orlando, FL., Nov. 2001.

[3] T. Keaton, S. M. Dominguez, and A. H. Sayed, "Snap&Tell: a vision-based wearable system to support web-on-the-world applications", *Proc. Digital Image Computing - Techniques and Applications Conference (DICTA)*, Melbourne, Australia, Jan. 2002.

[4] R. K. Mehra, "On the identification of variances and adaptive Kalman filtering", *IEEE Transactions on Automatic Control*, AC-15, pp. 175-183, 1970.

[5] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, NJ, 2000.

[6] A. H. Sayed, "A framework for state-space estimation with uncertain models", *IEEE Transactions on Automatic Control*, vol. 46, no. 7, pp.998-1013, July 2001.

[7] A. Wu, M. Shah, and N. Da Vitoria Lobo, "A virtual 3D blackboard: 3D finger tracking using a single camera", *Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 536-543, Grenoble, France, 2000.