Interacting Multiple Model (IMM) Kalman Filters for Robust High Speed Human Motion Tracking

Michael E. Farmer [§] , Rein-Lien Hsu*, and Anil K. Jain*
[§] Eaton Corporation
* Dept. of Computer Science & Engineering, Michigan State University
Email: MichaelEFarmer@eaton.com,{hsureinl, jain}@cse.msu.edu

Abstract

Accurate and robust tracking of humans is of growing interest in the image processing and computer vision communities. The ability of a vision system to track the subjects and accurately predict their future locations is critical to many surveillance and camera control applications. Further, an inference of the type of motion as well as to rapidly detect and switch between motion models is critical since in some applications the switching time between motion models can be extremely small. The Interacting Multiple Model (IMM) Kalman filter provides a powerful framework for performing the tracking of both the motion as well as the shape of these subjects. The tracking system utilizes a simple geometric shape primitive such as an ellipse to define a bounding extent of the subject. The utility of the IMM paradigm for rapid model switching and behaviour detection is shown for a passenger airbag suppression system in an automobile. The simplicity of the methods and the robustness of the underlying IMM filtering make the framework well suited for low-cost embedded real-time motion sequence analysis systems.

1. Introduction

Accurate and robust tracking of humans is of growing interest in the image processing and computer vision communities [1, 4, 5, 8]. The ability of a tracking system to follow the subjects and accurately predict their future locations is critical to many surveillance and camera control applications. The tracking, however, is complicated by the fact that humans do not exhibit one type of motion but rather tend to transition between a set of typical motions [5]. Each of these motions can be represented as a different motion model. There has been a considerable amount of work in remote target tracking applications using Interacting Multiple Model (IMM) filtering [1]. The utility of tracking these various model transitions and recognizing the motions that correspond to these models has also been considered [2, 5]. The transitions between these various models have been referred to as kinematic discontinuities that must be accounted for to reduce the prediction errors [4]. Deutscher et al. [4] have also shown that the Extended Kalman filter alone is unable to account for these transitions. This is clear since the Extended Kalman filter can only linearize non-linear dynamics models, but still relies on a fixed underlying process noise model so it cannot support these discontinuities [6, 7]. The IMM framework provides for multiple models by allowing distinct process noise models for each of the underlying motion models [1].

Traditionally, tracking of humans and identifying human dynamics involve low-level segmentation and exact modeling of limbs as well as torso to infer the dynamic events [2, 5]. Our approach utilizes a simpler and robust representation, which only focuses on the head and torso, and presumes that the limbs will follow the resulting motion. It is based on the observation that when trying to track an opponent, athletes in sports such as hockey, football, and soccer, watch the motion of the chest/torso area to prevent being fooled by intentional faking moves [12].

The aim of this paper is to show that with the IMM Kalman filter as the underlying framework and a very simple representation of the human form, considerable information can be derived regarding the sequence of human motions. The simplicity of the methods and the robustness of the underlying IMM filter make the framework well suited for embedded real-time low-cost surveillance systems, such as automotive airbag suppression systems.

2. Interacting Multiple Model Framework

Figure 1 shows the basic processing system for supporting the IMM image sequence analysis framework. The shape parameter extraction module determines the centroid ***(x,y)*** coordinates, the major and minor axes, and the in-plane rotation, θ. These serve as the input measurement vector for the subsequent Kalman tracking.

There are two paths for the track processing: one for tracking shape, and the other for tracking motion. Each of these will be addressed below.

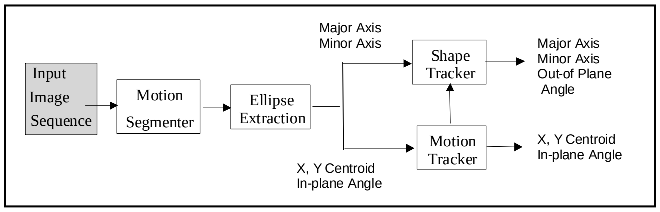

Figure 1. Processing flow for the tracking system.

Recall the basic model equations for the linear Kalman filter [6, 7]:

$$\begin{aligned} \boldsymbol{x}(k) &= \Phi(k-1) * \boldsymbol{x}(k-1) + \boldsymbol{v}(k-1) \\ \boldsymbol{z}(k) &= M(k) * \boldsymbol{x}(k) + \boldsymbol{w}(k), \end{aligned} \tag{1}$$

where Φ is the state transition matrix, $\boldsymbol{x}$ is the state vector, $\boldsymbol{v}$ is the process noise, $\boldsymbol{z}$ is the measurement value, $\boldsymbol{M}$ is the measurement matrix, and $\boldsymbol{w}$ is the measurement noise [6, 7]. The actual state estimation and prediction equations are given as:

$$\begin{aligned} \boldsymbol{x}(k \mid k-1) &= \Phi(k-1) * \boldsymbol{x}(k-1 \mid k-1) \\ \boldsymbol{x}(k \mid k) &= \boldsymbol{x}(k \mid k-1) + G(k) * residue(k) \\ residue(k) &= \boldsymbol{z}(k) - M(k) * \boldsymbol{x}(k \mid k-1) \\ S(k) &= M(k) P(k \mid k-1) M(k)^T + R(k), \end{aligned} \tag{2}$$

where *residue(k)* is the measurement residue, a Gaussian random variable with mean zero and covariance *S(k), and R(k)* is the covariance of the measurement noise ***w***.

The equations for the filter gain, ***G***, and the covariance matrix, ***P***, of the state prediction are then:

$$\begin{aligned} G(k) &= P(k \mid k-1) * M(k)^T * (M(k) * \\ &\qquad P(k|k-1) * M(k)^T + R(k))^{-1} \\ P(k \mid k-1) &= \Phi(k-1) * P(k-1 \mid k-1) * \\ &\qquad \Phi(k-1) + Q(k-1), \end{aligned} \tag{3}$$

where *Q(k)* is the covariance of the process noise ***v***, *M(k)* is the measurement matrix from Eq. (1), and $\Phi(k)$ is the state transition matrix from Eq. (1).

For the IMM implementation, there is effectively a complete set of Eqs. (2) and (3) for each model. The interaction between the models depends on the switching probabilities and the likelihoods of each of the models as shown in Fig. 2. The likelihoods are generated according to [1, 7]:

$$\begin{aligned} & x_{0m}(k-1 \mid k-1) = \sum_{s=1}^{N} x_s(k-1 \mid k-1) * f_{s|m}(k-1) \\ & f_{s|m}(k-1) = \frac{1}{\sum_{s=1}^{N} p(s \mid m) * f_s(k-1)} * p(s \mid m) * f_s(k-1) \\ & f_m(k) = \frac{1}{\sum_{s=1}^{N} L_s(k) * \sum_{t=1}^{N} p(t \mid s) * f_t(k-1)} * \\ & \qquad L_m(k) * \sum_{s=1}^{N} p(s \mid t) * f_s(k-1) \\ & L_m(k) = N[residue_m(k); 0, S_m(k)], \end{aligned} \tag{4}$$

where $f_{s|m}(k-1)$ is the probability of model ***s*** being correct at time *k-1*, given that model *m* is correct at time *k*, $f_m(k-1)$, $f_m(k)$ are the model probabilities at times *k-1* and *k*, respectively, $L_m(k)$ is the likelihood of the model *m* at time *k* based on the residue from the incoming measurement. Note that *N[x;μ,Σ]* stands for a normal distributiuon with an argument *x*, mean *μ*, and covariance *Σ*. The final output of the system is a combined state vector that is the sum of the state vectors for each of the modes weighted by their model probabilities.

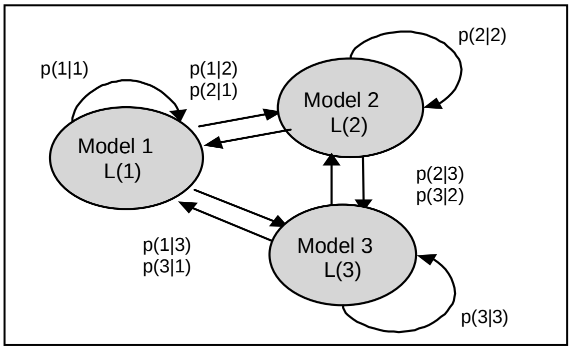

Figure 2. IMM model switching process.

The motion tracker processes the centroids of the subject as well as the in-image plane rotation angle θ as shown in Fig. 3. The motion of the subject can be represented by a set of models such as standing, walking and running. The key is to define the models as a set of states that can be represented by unique noise parameters. From the model probabilities we are able to handle dynamics that are between the two states by observing the relative probabilities. For example, a very fast gait would have a higher running probability than a slow walker but would still maintain a reasonable probability of walking. By characterizing the model probabilities for sets of subjects, it is possible to train a back-end classifier to recognize even more complicated dynamics by treating the underlying models as a basis function upon which other dynamics are built.

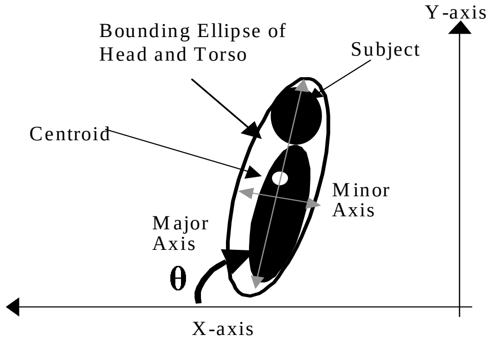

Figure 3. Geometry of tracking system.

The shape tracker is an Extended Kalman filter that is used to account for the interaction between the forward tilt angle, the lateral tilt angle and the corresponding deformations of the representative ellipse as the subject leans towards or away from the camera. This provides a crude estimate of the 3D orientation of the occupant.

3. Experimental Results

Recently, considerable attention has been paid in developing 'smart' airbags that can determine not only if they should be deployed in a crash event but also with what force they should be deployed [3, 9, 10, 11]. If the occupant (here, we are simulating the passenger and not the driver) is too close to the airbag, she may enter Automatic Suppression Zone (ASZ) in which case the airbag should not be deployed. Due to the rapidly changing dynamics, the tracking system must respond extremely quickly to changes in occupant motion.

The above framework was applied to this application. Other such airbag suppression systems require either multiple sensors for developing 3-D positional information [9, 10] or else require high-speed CMOS distance measuring devices [3]. The proposed IMM filter system has been shown to provide the ability to quickly detect the pre-crash event and successfully predict the future locations of the occupant to recommend airbag suppression. The tracking system performs the IMM-based Kalman filtering to accurately track the position of the occupant.

Our system uses a conventional 40 Hz CMOS imaging system, providing a 512x512 8-bit gray image (640x480 for human imagery). A robotic test fixture was created to propel a styrofoam dummy towards a vehicle instrument panel at a terminal speed of up to 9 miles per hour which is the limit of vehicle braking ability. The dummy in Fig. 4 (a) is moved in a constant acceleration mode that mimics the dynamics of a pre-crash braking event. Likewise, Fig. 5 shows the bounding ellipse for a human occupant in an image sequence containing 8 frames. For this application three modes of motion were used: (i) stationary, (ii) human stationary, and (iii) pre-crash brake. The three modes of motion span the entire dynamics space that the occupant can experience while in the vehicle. The stationary mode is when the person is sitting very still or is asleep. The human motion mode is when the occupant is moving about the seat in everyday movements such as opening gloveboxes, turning towards the driver or back seat, etc. The pre-crash braking mode representes the dynamics the occupant would experience when the vehicle is being stopped at or near the full force of the brakes.

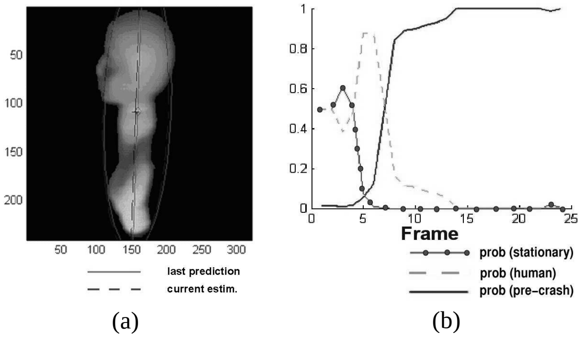

Figure 4. (a) Test dummy and bounding Ellipse; (b) model probabilities for pre-crash event.

The transition probabilities, *p(s|m)*, were determined a priori from analyzing videos of pre-crash braking events and simulating these events with crash dummy. Figure 4 (b) shows the history of the mode probabilities during the crash event. Note that for the first few frames, the dummy is stationary. After the initial frames, the tracker quickly begins transitioning to human motion as the dummy is initially moved in the crash sequence. After only two frames, the tracker determines the dynamics exceed what is expected for natural human motion and the system transitions to pre-crash mode. The system stays in pre-crash mode through the crash event and for a few moments afterwards. This hysteresis is intentionally implemented since the system is designed to require overwhelming evidence that the crash has ended before the filter is allowed to return to a low-gain state.

Figure 6 shows the history of the mode probabilities for a human occupant, shown in Fig. 5, moving in the seat at normal human motion rates and also sitting still for periods of time. Transitions from the human mode to the stationary mode appear in frames *14* and *39* in Fig. 6. during the first *40*-frame intervals. Note there is no hysteresis during the transitions to these states. The behavior of the transitions can be completely controlled through the underlying transition probabilities to allow the system to react in desired ways to particular motions.

Figure 7 shows the resultant tracking accuracies during the entire pre-crash braking event. Notice that the worst-case errors were on the order of less than ± 5 pixels during a sequence that includes a transition from fully stationary motion to a high speed pre-crash braking event.

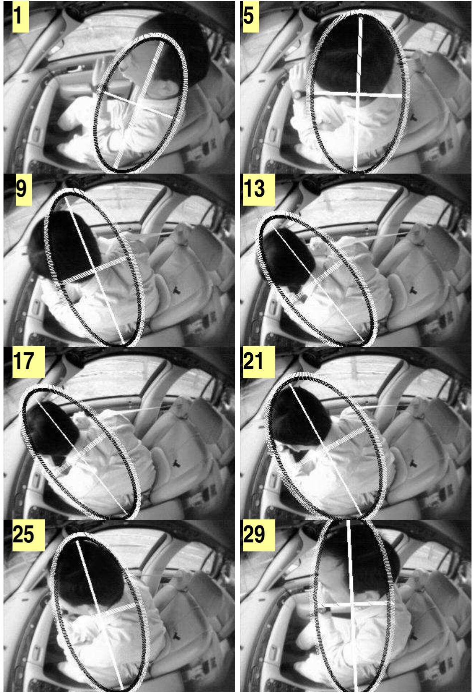

Figure 5. Human occupant and bounding ellipse shown in a sequence containing 8 frames.

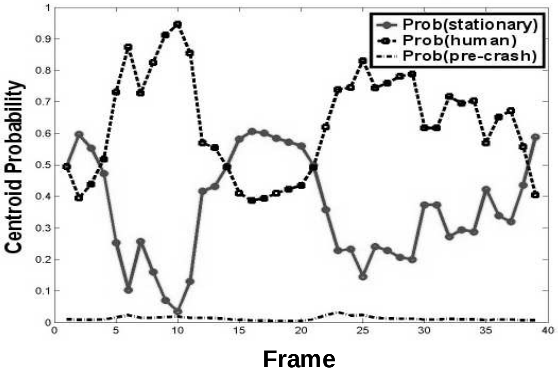

Figure 6. Model probabilities for human motion event.

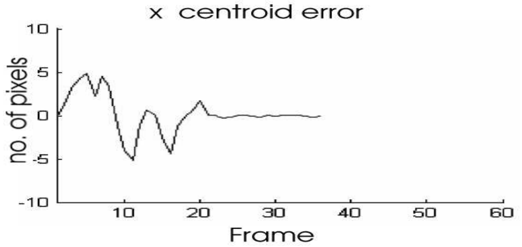

Figure 7. Plot of positional tracking accuracies (in pixels).

4. Summary and Conclusions

We have shown that the Interacting Multiple Model Kalman filter provides a robust framework for tracking human motion through a series of discontinuities. The mode switching capability allows the system to minimize tracking error by quickly adjusting the gain of the filter to reduce tracker latency. The mode switching also provides information on the type of motion the subject is employing. The tracking system is capable of minimal error even through unnatural events such as a high-speed pre-crash braking maneuver.

By focusing on the motion of the human torso and developing a simple representation of the torso, it is possible to develop a tracking system that can track through extremely sudden discontinuities. The simplified torso model allows us to ignore extraneous motions of the subject's limbs and other background motions and effectively track the human subject and determine their motion type. Through our ongoing research, we plan to show that this simplified human representation will also afford us the ability to infer 3-d subject pose and motion from 2-d image sequences in real-time without the need for complex point tracking and correspondence algorithms.

References

[1] H.A.P Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with markovian switching coefficients,'' *IEEE Trans. Automatic Control*, vol. 33 , no. 8, pp. 780-783, Aug. 1988.

[2] M. Isard and A. Blake, "A mixed-state condensation tracker with automatic model-switching,'' *Proc. ICCV*, pp. 107-112 , Jan. 1998.

[3] P. Mengel, G. Doemens, L. Listl, "Fast range imaging by CMOS sensor array through multiple double short time integration (MDSI),''*Proc. ICIP*, vol. II, pp. 169-172, Oct. 2001.

[4] J.B.N. Deutscher, B. Bascle, and A. Blake, "Tracking through singularities and discontinuities by random sampling, '' *Proc. ICCV*, vol. 2, pp. 1144-1149, 1999.

[5] C. Bregler, "Learning and recognizing human dynamics in video sequences,'' *Proc. IEEE Conf. CVPR*, June 1997.

[6] A. Gelb, Applied Optimal Estimation, *MIT Press*, 1974

[7] M. Pekkarinen, "Multiple model approaches to multisensor tracking,'' *Masters Thesis*, Tampere University of Technology, 1999.

[8] S. Wachter and H. Nagel, "Tracking of persons in monocular image sequences,'' *Proc. IEEE Workshop on Non-Rigid and Articulated Motion, CVPR*, 1997.

[9] A.P. Corrado, S. Decker, and P. Benbow, "Automotive occupant sensor system and method of operation by sensor fusion,'' *US Patent 5482314*.

[10] J.H. Semchena., E. Faigle, R. Thompson, J. Mazur, and C. Steffens Jr., "Apparatus and method for controlling an occupant restraint system,'' *US Patent 5531472*.

[11] J. Krumm and G. Kirk, "Video occupant detection for airbag deployment," *Proc. IEEE Workshop on Applications of Computer Vision*, pp. 30-35, Oct. 1998.

[12] S. Rossiter, *The NHL Way Hockey The Basics*, Greystone Books, 1996.