Pedestrian Localization by Appearance Matching and Multi-mode Filtering

Shunguang Wu, Mayank Bansal, Jayan Eledath
Sarnoff Corporation
201 Washington Rd, Princeton, NJ, USA
{swu,mbansal,jeledath}@sarnoff.com

Abstract— **This paper addresses the frame-to-frame data association and state estimation problems in localization of a pedestrian relative to a moving vehicle from a far infra-red video sequence. In a novel application of the hierarchical model-based motion estimation framework, we are able to solve the frame-to-frame data association problem as well as estimate a sub-pixel accurate height ratio for a pedestrian in two frames. To estimate the position and velocity of a pedestrian, instead of using a constant pedestrian height model, we propose a novel approach of using the interacting multiple-hypothesis-mode/height filtering algorithm. We present a method to calculate the probability of each mode from the estimated and measured pedestrian height ratios in images. These mode probabilities are then used to accurately estimate the pedestrian location by combining the mode based estimations. We demonstrate the effectiveness of our approach comparing it to a constant height model based approach on several IR sequences.**

Keywords—**Data association, pedestrian tracking, object scale measurement, multiple-hypothesis-mode filtering.**

I. INTRODUCTION

In recent years, there has been an increased use of visual sensors in automotive safety and convenience applications. One important safety application is to detect pedestrians[1] at night time. Visible-range cameras do not provide sufficient contrast to detect pedestrians well - a problem which is well handled by near and far infra-red (NIR,FIR) cameras. FIR cameras carry the advantage of target heat sensitivity without the need for active ambient illumination. The images of vehicles, pedestrians and animals are significantly enhanced and are clearly visible under otherwise poor visibility conditions. To keep the system cost low, current systems typically rely on a single FIR camera for pedestrian detection as well as range estimation. Accurately estimating the 3D location of the pedestrian relative to the moving vehicle is important for accurate warnings. This is a challenging problem as the system has to rely on the temporal tracking to estimate the location - both frame-to-frame data association as well as state-estimation filtering become important. In this paper, we will focus on the data-association and state-estimation aspects.

Gandhi et al.[2] have given a comprehensive survey of recent research on pedestrian collision avoidance systems. The paper reviews various approaches based on cues such as shape, motion, and stereo used for detecting pedestrians from visible as well as non-visible light sensors. Most of the approaches use image information for single-frame detection but not for associating these detections across frames. In [3], a Chamfer based coarse-to-fine strategy is applied to detect pedestrian candidates matching a predefined set of templates. However, the contour matching is not used for data-association between frames. Some amount of work has been done on tracking deformable objects in high-dimensional spaces using complex parameterized models of appearance and motion (e.g. [4]). These methods try to use filtering both for computing the object state as well as for refining the appearance model. This puts too much computational burden on the filter and does not use the appearance information directly across time.

In FIR imagery, the appearance of a pedestrian does not change much from frame-to-frame and it becomes possible to match the pedestrian not just with a pre-defined template set but also with the detection seen in the previous frame. This temporal image-based matching approach helps the tracker by a) reducing the state-space and hence the complexity of the filter required by not requiring an appearance model to be maintained by the filter, b) providing an alternate more robust means for data-association in case of missed-detections and c) explicitly estimating a sub-pixel object size ratio (which we call *scale*) in the image between two frames. In this paper, we describe a novel application of the hierarchical model-based motion estimation paradigm of [5] to match pedestrian appearance over time without explicitly modeling the pedestrian shape. The appearance matching is used both to resolve the frame-to-frame association of the detections as well as to estimate the scale across time which is an important component of the filter described in this paper.

Once the pedestrian bounding boxes are detected and temporal associations established, depending on the availability of the intrinsic and extrinsic camera parameters, a tracking process estimates the locations and velocities of pedestrians in either in image space or in host vehicle referenced 3D world coordinate system with particular data association techniques. Depending on the system and observation mod-

978-1-4244-3504-3/09/$25.00 ©2009 IEEE

eling approaches, the tracking algorithms could be Kalman filtering for linear systems [6], [7]; particle filtering [8], [9], [10], [11], and unscented Kalman filtering [12] for nonlinear systems; or adaptive interacting multiple models [13]. The localization process can project a ROI measurement to a 3D world frame by using the pin hole camera model. In this process, most researchers assume the height of the pedestrian (H) is constant for all kind of people, for example, $H = 1.65 \pm 0.1$m in [6] and $H = 1.8$m in [9]. However, the projected 3D distance error could be a significant factor because the difference between an assumed height and the real height can be as large as ±0.5 m.

To obtain a more accurate 3D localization, instead of using a constant H (one mode) for all pedestrians, this paper presents a multiple-hypothesis-mode filtering algorithm where each mode assumes a potential discrete height value for the pedestrian. Assuming that the pedestrian heights can be discretized into N bins or modes, N filters run in parallel as part of the filtering algorithm, and the probability of each filter is obtained by evaluating the likelihood value of an estimated pedestrian scale relative to the measured scale from the appearance matcher. The final pedestrian location can be obtained either by combining the N-mode estimations together or just choosing the one with the highest likelihood value. Experiments from several recorded IR sequences show the promised results of the proposed method.

In brief, the main contributions of this paper can be summarized as follows: (i) a novel application of hierarchical model-based motion estimation for temporal data-association, matching and scale estimation of detected pedestrians in FIR imagery, (ii) a novel approach of using multiple-hypothesis modes to solve the pedestrian localization problem and (iii) the concept of using the object size ratio (scale) between two frames to evaluate the likelihood value of each hypothesis mode.

The rest of the paper is organized as follows. Section II presents an overview of our system, and section III describes the pedestrian detection and appearance matching approaches. The single mode and multiple-hypothesis-mode filters are described in sections IV and V respectively. Experiment results are briefed in section VI, and conclusions are drawn in section VII.

II. Overview

Figure 1 presents an overview of our system. The inputs to the system are an FIR video stream and the vehicle speed and yaw rate measurements from the vehicle CAN bus. The *pedestrian detection* module detects candidate pedestrian ROIs in each frame and feeds them to the *appearance matching* module. This module takes in the current frame detections and the track predictions from the state-estimation filter and establishes appearance match based frame-to-frame associations. For each pedestrian ROI, it outputs the feet and head locations (from the ROI), and a scale estimate between the current and reference frames. The *ego-motion* module computes a pitch rate estimate using the image data. The *camera passed distance* filter uses the vehicle speed and yaw-rate measurements to estimate the distance the vehicle has traveled since the last frame. Finally, a *multi-mode-hypothesis* filter combines the pitch estimate, the vehicle passed distance estimate, the ROI feet and head locations and the scale estimates to compute the 3D location of the pedestrian relative to the camera and its velocity in the inertial frame.

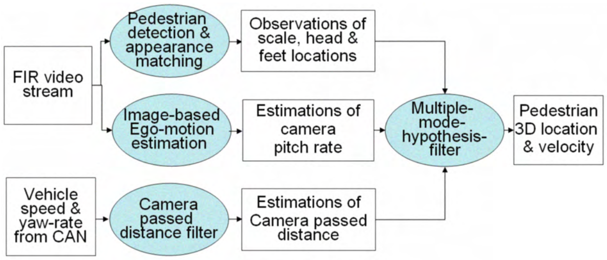
FIR video stream
Pedestrian detection & appearance matching
Observations of scale, head & feet locations
Image-based Ego-motion estimation
Estimations of camera pitch rate
Multiple-mode-hypothesis-filter
Pedestrian 3D location & velocity
Vehicle speed & yaw-rate from CAN
Camera passed distance filter
Estimations of Camera passed distance

Fig. 1. An overview of the system.

III. Pedestrian Detection and Appearance Matching

A FIR sensor images by passively sensing the heat-signature of the environment. Consequently, pedestrians and other warmer objects like vehicle undersides are imaged at brighter intensities. We follow the initial pedestrian detection approach in [14] by first selecting interesting regions by scanning for hot-spots in the image. The interesting regions found by the hot-spot detector provide seeds to an energy minimization based pedestrian model fitting algorithm which detects pedestrian aspect ROIs as initial detections. Thereafter, a multi-stage classifier is used to prune the initial detection set to give a set of candidates for tracking in successive frames.

Once a set of detections is available, a data-association step tries to associate new detections with any existing tracks. New tracks are started for detections never seen before and older tracks are terminated if no detections are seen for them contiguously for a few frames. For each existing track, an expected detection location is computed for this frame by projecting the world location predicted by the state-estimation filter (described in sections IV and V). An ROI overlap criterion is used to decide whether a new detection might possibly belong to this track. For all the detections that pass the overlap criterion, an image based appearance matching test is conducted between the ROI in the last frame and the candidates in this frame to decide the best matching candidate. The appearance matching test outputs a confidence measure which is used to decide the best matching candidate as well as to infer if there is a mis-detection. This helps with data-association in cluttered environments where pedestrians occlude each other (thus leading to a mis-detection) by avoiding association of pedestrians which are dissimilar in appearance but close in world locations. In addition, the matching step estimates a parametric transformation between the two detections which provides a scale estimate to the state-estimation filter. In case the matching step outputs a high confidence, the parametric transformation is also used to warp the tracked ROI to the current frame. This warped ROI is then used as the new measurement for this frame instead of the output from the single-frame detector. This reduces the dependence on the ROI detected by the single-frame detector which is typically very noisy.

Our appearance matching and scale-estimation algorithm is based on the hierarchical model-based motion estimation framework of [5]. Since detection of stable features over time is difficult in FIR imagery, it is ideal to use a dense direct-estimation framework like [5] to compute an appropriate motion model between frames. For this problem, we estimate a reduced affine motion model. This is because the local depth variation of the pedestrian is very small relative to its distance from the camera and thus, an affine motion-model is sufficient. Also, in the cases where the host-vehicle is directly approaching a pedestrian, there is sufficient change in the pedestrian size that a simple correlation based matching scheme (i.e. translation only model) would not work.

The appearance matching and scale estimation scheme is presented in Fig.2. The detected ROI in the last frame (time $t-1$) is expanded by 10% of the initial size and image pixels within this ROI serve as the reference image. Each ROI candidate close to the filter prediction in the image at time t is termed as the inspection image. The goal of the matching algorithm is to search for a transformation that relates the inspection image to the reference image. The direct-estimation method is applied in a coarse-to-fine manner on the laplacian image pyramids computed from the reference and inspection images. In this coarse-to-fine estimation framework, motion models with a lower number of parameters are estimated using images at coarser level and then used to seed estimation of more complex models using images at finer levels. This speeds up the estimation while also avoiding getting stuck in local minima. We estimate only a 2-parameter (t_x, t_y) translation motion model at the coarsest level and a 3-parameter (s, t_x, t_y) reduced affine motion model at finer levels. The parameter values are estimated with an accuracy of 0.1 of a pixel. The ROI detected at time $t-1$ is warped using the estimated transformation to compute a ROI at time t which is used as the measurement to the state-estimation filter. To keep the estimation errors from accumulating, in practice, we perform appearance matching between a reference frame $t-k$ ($k \geq 1$) in the past and the current frame t. The reference frame is kept fixed until the estimated scale becomes too large at which point it is reset to $t-1$.

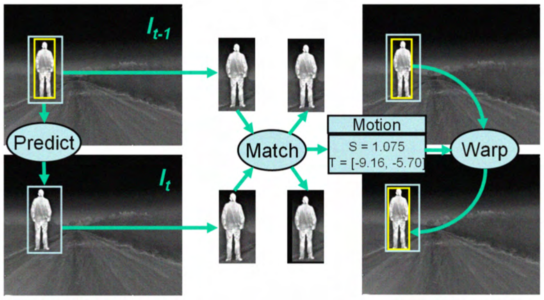

Fig. 2. Illustration of the appearance matching process to recover ROI at time t and the scale change between frames $t-1$ and t.

The appearance matching algorithm outlined above may fail in the special case where the pedestrian is moving laterally across the field-of-view due to significant leg motion. Thus, in our system, in general we estimate the transformation for the top and bottom halves of the ROI separately and then either output just the parameters from the top-half or re-estimate them for the whole ROI depending on whether the two sets of parameters are close (thus implying that the legs are in fact following the same motion parameters). We have seen a significant improvement in the lateral velocity estimation of pedestrians with the use of appearance matching. This is because it is difficult for a single-frame detector to output reliable bounding boxes around a pedestrian moving laterally while the appearance matcher estimates a much more accurate sub-pixel bounding box estimate by using information from the upper body.

Fig.6 shows an example of how the scale estimated from the appearance matching method is smoother compared to that estimated by just taking ratios of ROI heights in successive frames (height-ratio method). The zig-zag nature of the plot can be attributed to the varying separation between the current and the reference frames.

IV. SINGLE MODE FILTER

In the single mode filter, we assume the height of the pedestrian is known. As shown in Fig. 3, under the ground plane assumption, suppose a pedestrian is located at (X,Y) in a vehicle fixed coordinate system with a walking velocity of (v_x, v_y) in the inertial coordinate system at time t_k. The system state is defined as

$$\mathbf{x}_k = [X, Y, v_x, v_y, \theta]_k^T, \tag{1}$$

where θ is the pitch angle of the vehicle. Modeling the pitch angle as part of the state is important to be able to localize a pedestrian which is far away from the camera.

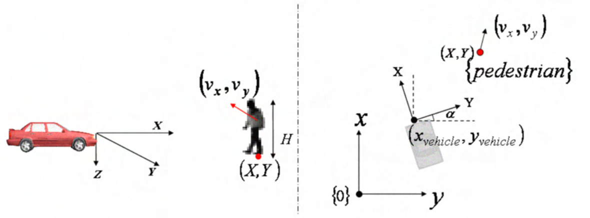

Fig. 3. Left: a 3D view of the camera coordinate system used in system modeling; Right: a bird-eye view of the system modeling coordinates.

Assuming that the pedestrian moves with a nearly constant velocity, its location in the camera reference frame can be modeled by a rotation (governed by the vehicle yaw angle change) and a translation (governed by the vehicle movement which shifts the pedestrian relative to the rotated frame). Similar to [12], from the geometric relationship shown in Fig.3, the kinematics equation between two consecutive frames k and $k+1$ can be written as

$$\mathbf{x}_{k+1} = A\mathbf{x}_k + B\mathbf{u}_k + \mathbf{w}_k, \tag{2}$$

where $\mathbf{w}_k$ is the kinematics modeling uncertainty which is assumed to be $\mathbf{w}_k \sim N(0, \mathbf{Q}_k)$ and the control input term, $\mathbf{u}_k = (v_k, \dot{\alpha}_k, \dot{\theta}_k)^T$, represents the speed, yaw and pitch rates of the camera. The speed and yaw rate are obtained directly from the vehicle CAN bus while the pitch rate is estimated by an image based ego-motion estimation module [15]. Let $T = t_{k+1} - t_k$, $\alpha = \dot{\alpha}T$, then the matrices A and B are expressed as,

$$A = \begin{bmatrix} c_\alpha & s_\alpha & Tc_\alpha & Ts_\alpha & 0 \\ -s_\alpha & c_\alpha & -Ts_\alpha & Tc_\alpha & 0 \\ 0 & 0 & c_\alpha & s_\alpha & 0 \\ 0 & 0 & -s_\alpha & c_\alpha & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} -Tc_\alpha & 0 & 0 \\ Ts_\alpha & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & T \end{bmatrix},$$

where $s_\alpha = \sin(\alpha)$, and $c_\alpha = \cos(\alpha)$.

The observation vector is defined as,

$$\mathbf{z}_k = [x_{feet}, y_{feet}, x_{head}, y_{head}]^T, \tag{3}$$

where (x_{feet}, y_{feet}) and (x_{head}, y_{head}) are the feet and head locations of the pedestrian in the image. In addition, assuming the camera projection parameters to be known, we have the following non-linear measurement equations,

$$\mathbf{z}_k = \mathbf{h}(\mathbf{x}_k) \overset{\triangle}{=} \begin{cases} h_1 = f_{px}\frac{Y}{X\cos(\theta) - Z_c\sin(\theta)} + \frac{I_w}{2} + n_1, \\ h_2 = f_{py}\frac{X\sin(\theta) + Z_c\cos(\theta)}{X\cos(\theta) - Z_c\sin(\theta)} + \frac{I_h}{2} + n_2, \\ h_3 = f_{px}\frac{Y}{X\cos(\theta) - (Z_c - H)\sin(\theta)} + \frac{I_w}{2} + n_3, \\ h_4 = f_{py}\frac{X\sin(\theta) + (Z_c - H)\cos(\theta)}{X\cos(\theta) - (Z_c - H)\sin(\theta)} + \frac{I_h}{2} + n_4, \end{cases} \tag{4}$$

where H is the world-height of the pedestrian, I_w and I_h are the width and height of the image respectively, f_{px} and f_{py} are the horizontal and vertical focal lengths of the camera respectively, Z_c is the height of the camera from the ground plane, and θ is the pitch angle of the camera relative to the ground plane. $\mathbf{n}_k = [n_1, n_2, n_3, n_4]_k^T$ is the observation noise term which is also modeled as a zero mean Gaussian with covariance $\mathbf{R}_k$.

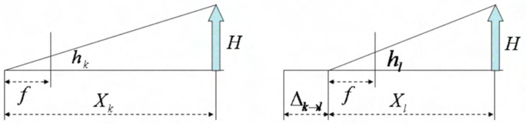

Fig. 4. This schematic illustrates the variation of an object's image height h as the camera (with focal-length f) moves a distance of $\Delta_{k\to l}$ between frames k and l.

Once we have the system and observation equations (2) and (4), the non-linear filtering algorithms can be applied to estimate the state of the system from its observation history. Particularly, in this work the extended Kalman filtering (EKF) is employed.

V. MULTI-MODE FILTER

The estimation results of the single mode filter (for results, refer to section VI) strongly rely on the *a priori* information of the pedestrian's height. Since the height variance of pedestrians can be as large as ±0.5m, our simulations show that a fixed height assumption can introduce unacceptable errors. For example, if we assume that $H = 1.6$m for a pedestrian whose actual height is $1.3(1.9)$m and is standing at a longitudinal distance of around 32 m, the estimated distance will be about 5 m more(less) than the ground truth.

To deal with this problem, the filter can be designed in the interacting multiple model (IMM) [16] framework. To do so, we first discretize the height range of pedestrians as $\{H_1, H_2, \cdots, H_N\}$, then for each height hypothesis, a corresponding single mode filter is applied and its likelihood value evaluated. The final estimated state can be obtained either by choosing the result from the largest likelihood value filter or by doing a weighted combination of individual filter results. The details of the multi-mode filter is presented in the following subsections.

A. Mode Likelihood Value Evaluation

To evaluate the likelihood value of a single mode filter corresponding to a particular pedestrian height, we derive the formula to estimate the scale of a pedestrian for a given camera passed distance first. As shown in Fig. 4, let H be the height of a pedestrian in the 3D world, X_k be his distance from the camera (as defined in (1)), and h_k be his height in the image (in pixels) at frame k. Let $\Delta_{k\to l}$ denote the camera passed distance from frames k to l ($l > k$). From $h_k X_k = h_l X_l$ one has,

$$s_{k\to l} \overset{\triangle}{=} \frac{h_l}{h_k} = \frac{X_k}{X_l} = \frac{X_k}{X_k - \Delta_{k\to l}}, \tag{5}$$

where $s_{k\to l}$ represents the scale of the pedestrian in the image between frames k and l.

Equation (5) shows that the scale is determined by the distance between the camera and the pedestrian at frame

k and the camera passed distance between frames k and l. In addition, given the estimated camera to pedestrian distance and its variance pair $(\hat{X}_k, \sigma^2_{\hat{X}_k})$, and the camera passed distance and is variance pair$(\hat{d}_k, \sigma^2_{\hat{d}_k})$, we can compute the estimated scale and its variance as follows:

$$\hat{s}_{k\to l} = \frac{\hat{X}_k}{\hat{X}_k - \Delta}, \sigma^2_{\hat{s}} = c_1^2\sigma^2_\Delta + c_2^2\sigma^2_{\hat{X}_k}, \tag{6}$$

where $\sigma^2_\Delta = \sigma^2_{\hat{d}_k} + \sigma^2_{\hat{d}_l}$, $c_1 = \frac{\hat{X}_k}{(\hat{X}_k - \Delta)^2}$, $c_2 = -\frac{\Delta}{(\hat{X}_k - \Delta)^2}$, and $\Delta = \hat{d}_l - \hat{d}_k$.

Note that the scale estimated from (6) is implicitly dependent on the height of the pedestrian hypothesized by the single-mode filter. Let $(s_{k\to l}, \sigma^2_s)$ be the actual scale and its variance estimated by the appearance matching algorithm (described in section III). Then, the likelihood of this mode can be represented as

$$\Lambda_{k\to l} = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\hat{s}_{k\to l} - s_{k\to l})^2}{2\sigma^2}), \tag{7}$$

where $\sigma^2 = \sigma^2_s + \sigma^2_{\hat{s}}$.

B. Camera Passed Distance Estimation

The camera passed distance $\Delta_{k\to l}$ can be estimated by filtering the speed and yaw rate data obtained from the vehicle CAN bus. The state vector of the camera passed distance filter is defined as:

$$\mathbf{x}_k = [d, v, \dot{\alpha}]^T_k, \tag{8}$$

where d, v, and $\dot{\alpha}$ are the camera passed distance, the vehicle speed and the vehicle yaw rate, respectively. The kinematics equation of this filter can be modeled as:

$$\begin{aligned} d_{k+1} &= d_k + T\cos(\dot{\alpha}_k T) + w_1(k), \\ v_{k+1} &= v_k + w_2(k), \\ \dot{\alpha}_{k+1} &= \dot{\alpha}_k + w_3(k), \end{aligned} \tag{9}$$

where $\mathbf{w}_k = (w_1, w_2, w_3)^T$ is the uncertainty term which is assumed to be zero mean Gaussian with constant covariance.

The measurement vector and its equation are respectively defined as:

$$\mathbf{z}_k = (v, \dot{\alpha})^T_k, \text{and } \mathbf{z}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{n}_k, \tag{10}$$

where $\mathbf{H}_k = (0\ 1\ 0;\ 0\ 0\ 1)$ and $\mathbf{n}_k \sim N(0, \mathbf{R}_k)$ with $\mathbf{R}_k = diag(\sigma_v(k), \sigma_{\dot{\alpha}}(k))$.

Due to the non-linearity of the kinematics equation, the camera-passed distance filter is implemented as an EKF.

C. Interacting Multi-Mode Implementation

The multiple EKFs which correspond to multiple pedestrian height hypothesis are integrated under the interacting multiple mode (IMM) framework [16]. As shown in Fig. 5, assuming that we have N discrete height hypotheses, the implementation of IMM filter includes the following steps:

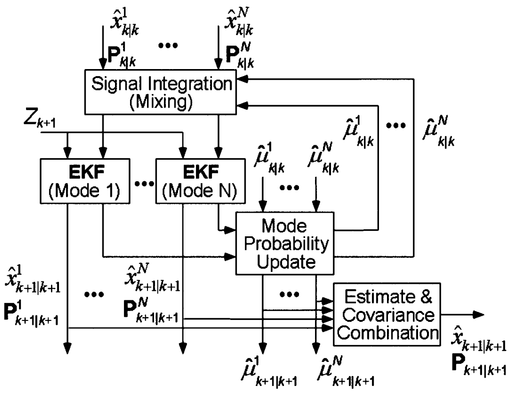

Fig. 5. The data flow of the IMM algorithm.

(i) *Initialization.* Given the camera intrinsic and extrinsic parameters, the height of the pedestrian H, and its head (x_{head}, y_{head}) and feet (x_{feet}, y_{feet}) locations in the image, its initial state vector, $\mathbf{x}_{0|0} = [X_0, Y_0, v_{x_0}, v_{y_0}, \theta_0]^T$, is given by

$$\begin{aligned} X_0 &= Hf_{py}/(y_{feet} - y_{head}), \\ Y_0 &= X(x_{feet} - 0.5I_w)/f_{px}, \\ v_{x_0} &= v_{y_0} = 0, \\ \theta_0 &= \text{atan2}[-\tfrac{X(y_{feet} - 0.5I_h) - Z_c f_{py}}{(y_{feet} - 0.5I_h)Z_c/f_{py} - X}, f_{py}], \end{aligned} \tag{11}$$

where I_w, I_h, f_{px}, f_{py}, and Z_c are as in (4). The corresponding covariance matrix can be set from the information of the observation covariance, the Jacobian of (11) and independent initial parameters for velocity elements.

(ii) *Mode mixing.* In the mode mixing block, first we need to mix the mode probabilities i.e.,

$$\mu^{j|i}_{k|k} = \frac{1}{\bar{c}_i} p_{ji}\mu^j_k, \quad \bar{c}_i = \sum_{j=1}^{N} p_{ji}\mu^j_k, \quad (i, j = 1, \ldots, N), \tag{12}$$

where p_{ji} is an element of the mode transition probability matrix, and μ^j_k is the probability of model j at time t_k. Then the mixed states and their covariances are calculated by

$$\bar{\mathbf{x}}^i_{k|k} = \sum_{j=1}^{N} \hat{\mathbf{x}}^j_{k|k}\mu^{j|i}_{k|k}, \quad \bar{\mathbf{P}}^i_{k|k} = \sum_{j=1}^{N} \mu^{j|i}_{k|k}[\mathbf{P}^j_{k|k} + \tilde{\mathbf{x}}^{ji}_{k|k}(\tilde{\mathbf{x}}^{ji}_{k|k})^T], \tag{13}$$

respectively, where $\tilde{\mathbf{x}}^{ji}_{k|k} = \hat{\mathbf{x}}^j_{k|k} - \bar{\mathbf{x}}^i_{k|k}$, and $i = 1, \ldots, N$.

(iii) *Mode updation and likelihood value evaluation.* By using the corresponding mixed results (13) as inputs, the j^{th} mode EKF is updated by the observation $\mathbf{z}_{k+1}$ to yield its estimate at t_{k+1}, e.g., $(\hat{\mathbf{x}}^j_{k+1|k+1}, P^j_{k+1|k+1})$. The corresponding likelihood value can be calculated by (7) and is denoted by Λ^j_{k+1}.

(iv) *Model probability calculation.* The weight of model j can be updated by

$$\mu^j_{k+1} = \frac{1}{c}\Lambda^j_{k+1}\bar{c}_j \qquad (j = 1, \cdots, N), \tag{14}$$

where $c = \sum_{j=1}^{N} \Lambda^j_{k+1}\bar{c}_j$, and $\bar{c}_j$ is defined in (12).

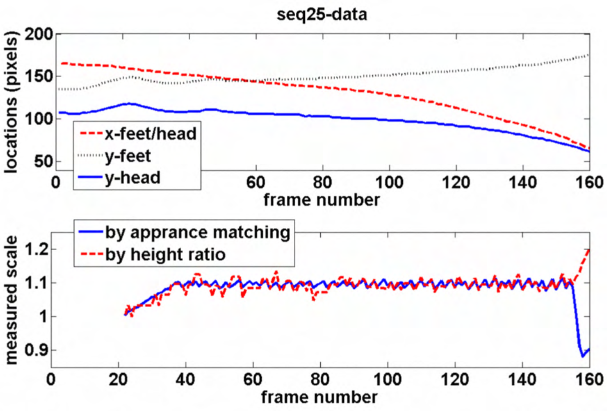

Fig. 6. The original observation data from sequence number 25. Top: measured head and feet locations in the image; Bottom: measured scales by appearance matching and height ratio methods.

(v) *State combination*. The final estimated state and its covariance are respectively computed as

$$\hat{\mathbf{x}}_{k+1|k+1} = \sum_{j=1}^{N} \hat{\mathbf{x}}^{j}_{k+1|k+1}\mu^{j}_{k+1}, \tag{15}$$

$$\mathbf{P}_{k+1|k+1} = \sum_{j=1}^{N} \mu^{j}_{k+1}(\mathbf{P}^{j}_{k+1|k+1} + \tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^T), \tag{16}$$

where $\tilde{\mathbf{x}}_j = \hat{\mathbf{x}}^{j}_{k+1|k+1} - \hat{\mathbf{x}}_{k+1|k+1}$.

VI. RESULTS

The proposed algorithm was tested on seven FIR sequences with known pedestrian heights (1.66m, 1.70m, 1.86m and 1.92m). These sequences cover a wide variety of driving and pedestrian scenarios. These include driving along straight roads as well as turns with multiple pedestrians standing still (in front, to the left and to the right of the vehicle), walking towards the vehicle, or walking laterally across the vehicle path etc.

The images are captured at 30 Hz at a resolution of 324×256 by a camera with horizontal and vertical focal-lengths $f_{px} = 498.5847$ and $f_{py} = 505.0273$ respectively, mounted on the front bumper of the car at a height of 0.65 m above the ground. For the single mode EKF, the system covariance matrix is set as $\mathbf{Q} = diag(3.0864 \times 10^{-7}, 3.0864 \times 10^{-7}, 0.00083333, 0.00083333, 0.01)$, its cross terms for velocity are $Q(3,1)$=$Q(1,3)$ = $Q(4,2)$ = $Q(2,4)$ = 1.3889×10^{-5}, and all the other elements are zero; the observation covariance matrix is set as $R = diag(5,5,5,5)$. For the vehicle passed distance filter, its system uncertainty variances of speed and yaw rate are $\sigma_s^2 = 1(m^2/s^2)$, $\sigma_{\dot{\theta}} = 0.01(rad^2/s^2)$. The measurement covariance matrix is set to $R = diag(1, 0.01)$.

Fig.6 displays the original measurements of the feet/head locations of a pedestrian in the image and its scale between the current frame and a reference frame for one sequence. The foot/head locations are obtained from the detection and appearance matching algorithms described in section III. The scales are computed by both appearance matching and the height ratio features. Comparing with the height ratio feature, the appearance matching gives a more smooth result, however it may fail whenever the appearance matching is not reliable (e.g., in the frame range of $[155, 160]$ at the bottom plot of Fig. 6). Hence, the scales from height ratio are used by the filter whenever the confidence value from the appearance matcher is low.

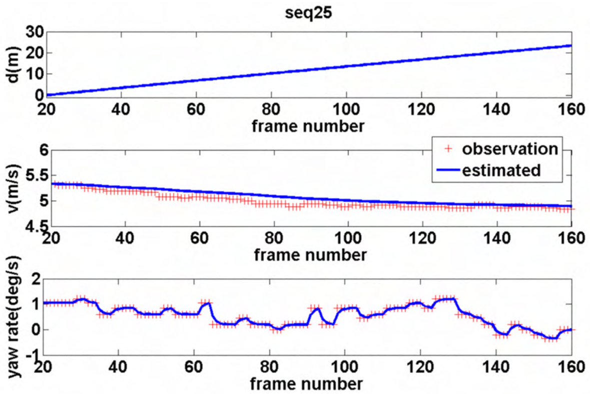

Fig. 7. The observation and estimated results of the passed distance filter.

The measured speed and yaw-rate results for the sequence are shown in *red pluses* in Fig. 7. The estimated results of the vehicle passed distance, its speed and yaw rate are shown in the same figure by *blue solid lines*.

Fig.8 shows the intermediate data from both IMM and single mode EKF filters. The left sub-plot is a comparison of the measured and estimated scales from different modes. Since the ground truth of the pedestrian height is 1.92 m, the estimated scales from the mode $H = 1.9$ m are almost overlapped on the observed ones. This verifies the correctness of our scale estimation approach. The middle and right sub-plots display the likelihood values used and mode probabilities estimated from the IMM algorithm. They indicate that the mode $H = 1.90$ m has the largest likelihood values as well as the highest mode probabilities, since it is the closest one to the ground truth height.

Finally, in Fig. 9, we plot the estimated state vector from the IMM filter and four single mode filters. The single mode filters include three from the modes used in the IMM, and one from the mode with ground truth height ($H = 1.92$ m). It is clear from these plots that the state estimated by the IMM filter is very close to the estimate from a single mode filter with known ground truth height.

VII. CONCLUSIONS

In summary, we propose a novel algorithm to do temporal data-association, matching and scale estimation from detected pedestrian ROIs in FIR image sequences by applying the hierarchical model-based motion estimation. In order to accurately localize a pedestrian in the vehicle fixed reference

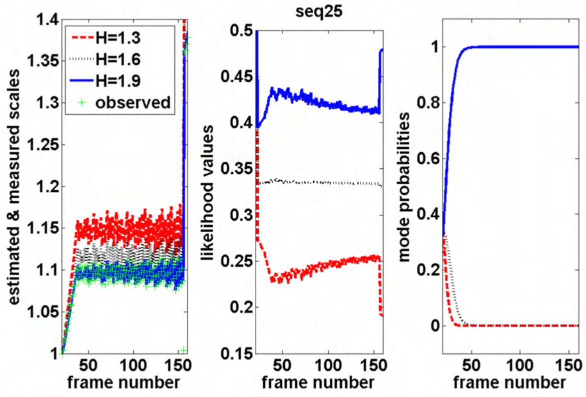

Fig. 8. Left: A comparison of the used observation scale and the estimated scales from different pedestrian height hypotheses; Middle: the normalized likelihood values from different modes; Right: the evolution of mode probabilities in IMM algorithm. This figure is best viewed in color.

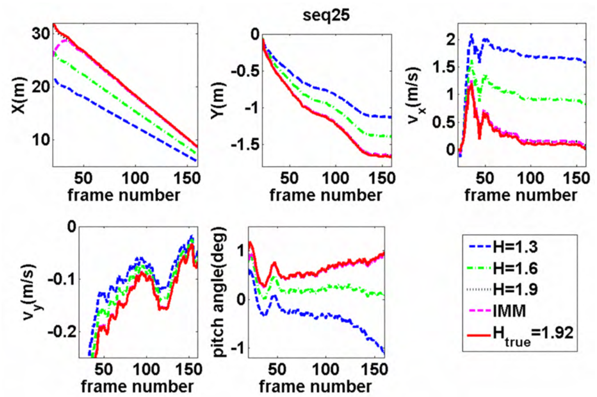

Fig. 9. A comparison of the elements of the state vector from single mode EKFs and the IMM filter, where three of the single mode filters use the heights used in the IMM and one uses the ground truth height. This figure is best viewed in color.

coordinate system, instead of using a constant pedestrian height assumption, we propose to use the multi-hypothesis-mode filtering. The likelihood value of each hypothesis mode can be evaluated using the scale computed from the appearance matching process and estimates of the vehicle passed distance and the longitudinal distance of the pedestrian. Additionally, the idea of multi-hypothesis-mode filtering is implemented under the framework of an IMM algorithm. The results from seven FIR video sequences demonstrate the success of the proposed algorithm.

Using an IMM kind of framework to implement the multiple-hypothesis-pedestrian-height filtering is one approach. Another approach is to use a particle filter. To do so, one can put the pedestrian height in the state vector defined by (1), and use discretized heights to generate different hypotheses. The likelihood of each hypothesis can still be evaluated using the method described in section V.C. The only concern would be the heavier computational load relative to the proposed method.

Finally, we should point out that besides the noise characteristics of the other observation data, the estimation errors of the proposed method are highly dependent on the accuracy of scale values measured from the appearance matching algorithm and the CAN bus data quality. Quantitatively evaluating the effects of these two factors will be one of our future tasks.

VIII. ACKNOWLEDGEMENTS

This work was partially supported by Autoliv Electronics AB, Linköping, Sweden. The authors would like to thank Vincent Mathevon, Elisabeth Ågren and Stefan B. Johansson for helpful discussions and data collection.

REFERENCES

[1] A. Shashua, Y. Gdalyahu, and G. Hayun, “Pedestrian detection for driving assistance systems: Single-frame classification and system level performance,” in *IEEE Intelligent Vehicles Symposium*, 2004.

[2] T. Gandhi and M. Trivedi, “Pedestrian collision avoidance systems: a survey of computer vision based recent studies,” in *IEEE Intelligent Transportation Systems Conference*, 2006, pp. 976–981.

[3] D. Gavrila, J. Giebel, M. Perception, D. Res, and G. Ulm, “Shape-based pedestrian detection and tracking,” in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, 2002.

[4] A. Baumberg and D. Hogg, “An efficient method for contour tracking using active shape models,” in *Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994, pp. 194–199.

[5] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, “Hierarchical model-based motion estimation,” in *European Conference on Computer Vision*. Springer, 1992.

[6] M. Bertozzi, A. Broggi, A. Fascioli, and A. Tibaldi, “Pedestrian localization and tracking system with kalman filtering,” in *Proc. IEEE Intelligent Vehicles Symposium, Parma, Italy*, 2004, pp. 584–589.

[7] F. Xu, X. Liu, and K. Fujimura, “Pedestrian detection and tracking with night vision,” *IEEE Trans. on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63– 71, 2005.

[8] K. Okuma, A. Taleghani, N. Freitas, and J. J. Little, “Boosted particle filter: Multitarget detection and tracking,” in *The 8th European Conference on Computer Vision (ECCV)*, 2004, pp. 28–39.

[9] R. Arndt, R. Schweiger, W. Ritter, D. Paulus, and O. Lohlein, “Detection and tracking of multiple pedestrians in automotive applications,” in *IEEE Intelligent Vehicles Symposium*, 2007, pp. 13–18.

[10] S. Gidel, C. Blanc, T. Chateau, P. Checchin, and L. Trassoudaine, “Nonparametric data association for particle filter based multi-object tracking: application to multi-pedestrian tracking,” in *IEEE Intelligent Vehicles Symposium*, 2008, pp. 73–78.

[11] S. Munder, C. Schnrr, and D. M. Gavrila, “Pedestrian detection and tracking using a mixture of view-based shapetexture models,” *IEEE Trans. on Intelligent Transportation Systems*, 2008.

[12] M. Meuter, U. Iurgel, S.-B. Park, and A. Kummert, “The unscented kalman filter for pedestrian tracking from a moving host,” in *IEEE Intelligent Vehicles Symposium*, 2008, pp. 37–42.

[13] J. Burlet, O. Aycard, A. Spalanzani, and C. Laugier, “Pedestrian tracking in car parks : an adaptive interacting multiple models based filtering method,” in *IEEE Intelligent Transportation Systems Conference*, 2006, pp. 462–467.

[14] J. Kallhammer, D. Eniksson, G. Granlund, M. Felsberg, A. Moe, B. Johansson, J. Wiklund, and P. Forssen, “Near Zone Pedestrian Detection using a Low-Resolution FIR Sensor,” in *IEEE Intelligent Vehicles Symposium*, 2007, pp. 339–345.

[15] S. Jung, J. Eledath, S. Johansson, and V. Mathevon, “Egomotion Estimation in Monocular Infra-red Image Sequence for Night Vision Applications,” in *Proceedings of the Eighth IEEE Workshop on Applications of Computer Vision*, 2007.

[16] Y. Bar-Shalom and W. D. Blair, *Multitarget-Multisensor Tracking: Applications and Advances, III*. Boston, MA: Artech House, 2000.