Pedestrian Association and Localization in Monocular FIR Video Sequence

Mayank Bansal, Shunguang Wu, Jayan Eledath
Sarnoff Corporation
201 Washington Rd, Princeton, NJ, USA
{mbansal,swu,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 monocular far infra-red video sequence. Using a novel application of the hierarchical model-based motion estimation framework, we are able to use the image appearance information to solve the frame-to-frame data association problem and estimate a sub-pixel accurate height ratio for a pedestrian in two frames. Then, to localize the pedestrian, we propose a novel approach of using the pedestrian height ratio estimates to guide an interacting multiple-hypothesis-mode/height filtering algorithm instead of using a constant pedestrian height model. Experiments on several IR sequences demonstrate that this approach achieves results comparable to those from a known pedestrian height thus avoiding errors from a constant height model based approach.

1. 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[17] 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. 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.

In FIR imagery, the appearance of a pedestrian does not change much from frame-to-frame and it becomes possible to match a pedestrian across time. 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 an application of the hierarchical model-based motion estimation paradigm of [4] to match pedestrian appearance over time without explicitly modeling the pedestrian shape. The appearance matching is used, first, to resolve the frame-to-frame association of the detections and then, to estimate the scale across time which allows a multiple-hypothesis-mode filtering algorithm to be employed for the state estimate phase.

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 and runs as a separate filter. 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 mode estimations together or just choosing the one with the highest likelihood value.

Related Work. Gandhi et al.[8] have given a comprehensive survey of recent research on pedestrian collision avoidance systems. The paper reviews various approaches

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

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 [9], 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. [3]). 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.

Once the pedestrian bounding boxes are detected and temporal associations established, a tracking process estimates the locations and velocities of pedestrians either in image space or in host vehicle referenced 3D world coordinate system. Depending on the system and observation modeling approaches, the tracking algorithms could be Kalman filtering for linear systems [5, 18], particle filtering [16, 1, 10, 15] and unscented Kalman filtering [13] for nonlinear systems, or adaptive interacting multiple models [6]. The localization process can project a ROI measurement to a 3D world frame by using the camera calibration information. In this process, most researchers assume the height of the pedestrian (H) is constant (e.g. $H = 1.65 \pm 0.1$m in [5] and $H = 1.8$m in [1]). However, the projected 3D distance error can be significant because the difference between an assumed height and the real height can be as large as ± 0.5 m.

In brief, the main contribution of this paper is the novel combination of using an object size ratio computed from image appearance information to guide a multi-hypothesis mode filtering algorithm for accurate pedestrian localization. The image appearance information is exploited using the hierarchical model-based motion estimation framework. We also show how the appearance matching in this framework can be used for temporal data association.

The rest of the paper is organized as follows. Section 2 presents an overview of our system, and section 3 describes the data-association and appearance matching approaches. Pedestrian Tracking using single and multiple-hypothesis-mode filters is described in section 4. Experimental results are briefed in section 5, and conclusions are drawn in section 6.

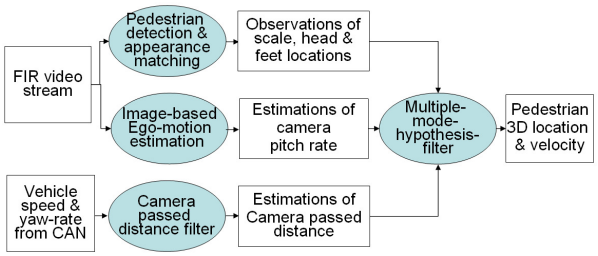

Figure 1. An overview of the system.

2. 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 [11] 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.

3. Data Association and Appearance Matching

In this paper, we will focus on the data-association and tracking aspects and assume that a separate pedestrian detector module is available. In our system, we follow the initial pedestrian detection approach in [12] 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.

3.1. Data Association

Once a set of detections is available, a data-association step tries to associate new detections with any existing

Figure 2. Example of data association by appearance matching for a sequence from the Terravic Motion IR Database[14]. The first row shows the original sequence with pedestrians labeled. The second and third rows show four different data-association configurations between the first and the fifth frames. For each configuration, we compute an image difference directly (third row) and after appearance matching the top part of the pedestrians (second row). The top-half of the error images show how the appearance matching makes it more robust to compare valid (A4-A0,B4-B0) and invalid (A4-B0,B4-A0) configurations by eliminating differences arising out of small scale changes even for valid configurations.

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 section 4). 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. An example is shown in Fig.2.

The matching step also 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.

3.2. Appearance Matching

Our appearance matching and scale-estimation algorithm is based on the hierarchical model-based motion estimation framework of [4]. Since detection of stable features over time is difficult in FIR imagery, it is ideal to use a dense direct-estimation framework like [4] to compute an appropriate motion model between frames. For this problem, we estimate a reduced affine motion model (translation + isotropic scaling). 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.3. The detected ROI in the last frame (time $t-1$) is expanded by an amount dependent on the vehicle speed (typically 10% of the previous ROI 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. Once the transformation is estimated, the ROI detected at time $t-1$ is warped using this transformation to compute a ROI at time t which is used as the measurement to the state-estimation filter.

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. The number of pyramid levels is adaptively chosen to ensure that the smallest pyramid level is bigger than a minimum size.

Fig.4 illustrates this coarse-to-fine approach for registering two pedestrian ROIs detected at different times. L_0, L_1 and L_2 are the successive levels of the laplacian pyramid computed as in [7]. I_x and I_y are the image gradients computed at each level of the reference image pyramid and I_t is the difference between the laplacian images at corresponding levels of the reference and inspection pyramids. Note

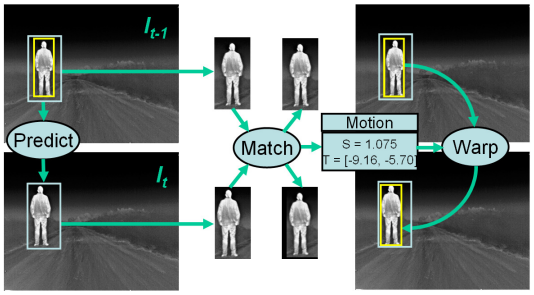

Figure 3. Illustration of the appearance matching process to recover ROI at time t and scale change between $t-1$ and t.

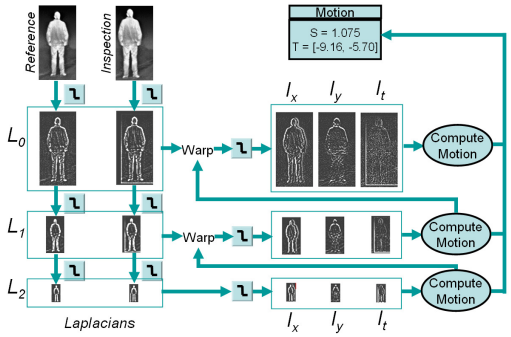

Figure 4. Illustration of the direct coarse-to-fine motion estimation process between two sample pedestrian ROIs.

that the gradient maps have been contrast stretched for ease of visualization. A translation only ($T = (t_x, t_y)$) motion model is first estimated between images at level L_2. The images at level L_1 are warped using the motion estimated at L_2 and then a translation and scale ($T+S$) motion model is estimated between the warped images. Similar process is repeated for L_0 where the final motion is computed. Note that at each level the image I_t depicts the amount of residual motion between the images at that level before the motion estimation step. It is clear that the coarse-to-fine strategy successively reduces the registration error between the candidate images thus progressing towards a finer motion estimate from level L_2 to L_0. The final residual gives a measure of the confidence in the computed transformation which is used by the data-association step.

Note that it is very important to get an accurate subpixel scale estimate for our problem and our registration strategy manages to achieve that quite well. The parameter values are estimated with an accuracy of 0.1 of a pixel. To keep the estimation errors from accumulating, in practice, the separation between the reference and the current frames is chosen adaptively.

Reference frame selection. Experiments with the appearance matching algorithm indicate that the estimated scale is most accurate when the actual scale is within a range of $[s_{lo}, s_{hi}]$. This range was empirically determined to be $[1.02, 1.04]$. Thus, to ensure that we operate in this range of scales, we adaptivly determine the separation k which will be used in the next frame using the scale estimated in the current frame. If the scale estimated in the current frame s_t (using a frame separation k_t) does not belong in the required range, we search for a separation k_{t+1} which will bring the estimated scale in the next frame in the required range i.e. we estimate the value $k_{t+1} \in [1, k_{max}]$ such that $s_t^{\frac{k_{t+1}}{k_t}} \in [s_{lo}, s_{hi}]$. The maximum separation k_{max} is dependent on the minimum scale that we need to estimate. In our experiments, we kept it at 5 frames. To achieve dynamic selection of a reference frame, a buffer of last k_{max} image frames is maintained along with an ROI list of the detections in each of those frames.

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). Fig.5 shows an example of this registration scheme in action for a pedestrian moving laterally as well as longitudinally across the field-of-view leading to significant scale change together with significant leg motion. 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.9 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. For this plot, the scale range has been constrained to a range different from $[1.02, 1.04]$.

4. Pedestrian Tracking

Once the pedestrian bounding boxes in the image space are detected and temporal associations established, a track-

Figure 5. Example of appearance matching for a difficult sequence from the Terravic Motion IR Database[14]. Every 10^{th} frame from the original sequence (top row) and the corresponding frames registered to the first frame (bottom row). The green line shows how well the scales of all the frames match the first frame after registration. Note how the registration can robustly handle scale change, motion in the legs and even a slight rotation of the person's head.

ing process estimates the locations and velocities of pedestrians in a host vehicle referenced 3D world coordinate system by using the camera calibration information. In this process, most researchers assume the height of the pedestrian (H) is constant (e.g. $H = 1.65 \pm 0.1$m in [5] and $H = 1.8$m in [1]). However, the projected 3D distance error can be significant 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, we will use 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 quantized 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 obtained from the appearance matcher. The final pedestrian location can be estimated in the sense of an Interacting Multi-Model (IMM) algorithm [2]. The following subsections present single filter modeling, the mode likelihood value evaluation and the implementation of the IMM algorithm.

4.1. Single Mode Filter

In the single mode filter, we assume the height of the pedestrian is known. As shown in Fig. 6, 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

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

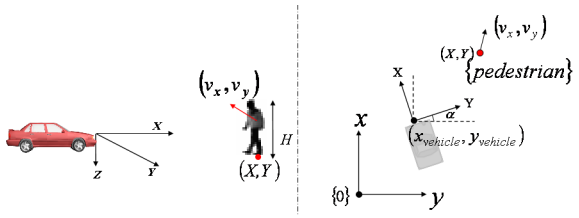

Figure 6. Left: a 3D view of the camera coordinate system used in system modeling; Right: a bird-eye view of the system modeling coordinates.

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.

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 [13], from the geometric relationship shown in Fig.6, 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 [11]. Let $T = t_{k+1} - t_k$, $\alpha = \dot{\alpha}T$, then the matrix A is expressed as,

$$A = \begin{bmatrix} R_\alpha & TR_\alpha & O_{2\times 1} \\ O_{2\times 2} & R_\alpha & O_{2\times 1} \\ O_{1\times 2} & O_{1\times 2} & 1 \end{bmatrix},$$

where $R_\alpha = [\cos(\alpha), \sin(\alpha); -\sin(\alpha), \cos\alpha]$, and $O_{m\times n}$ is an $m \times n$ zero matrix. B is a 5×3 matrix with $B(1,1) = -T\cos\alpha$, $B(2,1) = T\sin\alpha$, $B(5,3) = T$, and all other elements zero.

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

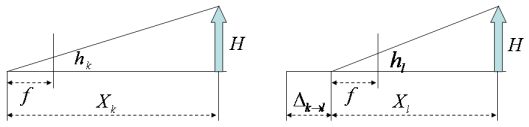

Figure 7. This schematic illustrates the variation of an object's image height h as the camera (with focal-length f) moves a distance of $\Delta_{k\to l}$ between frames k and l.

have the following non-linear measurement equations,

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

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 $\mathbf{R}_k$.

With the system and observation equations in (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 we found the extended Kalman filter (EKF) to be adequate.

4.2. 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. 7, 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,

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

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

4.3. 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:

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

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:

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

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

The camera-passed distance filter is implemented as an EKF because of the nonlinearity of (9).

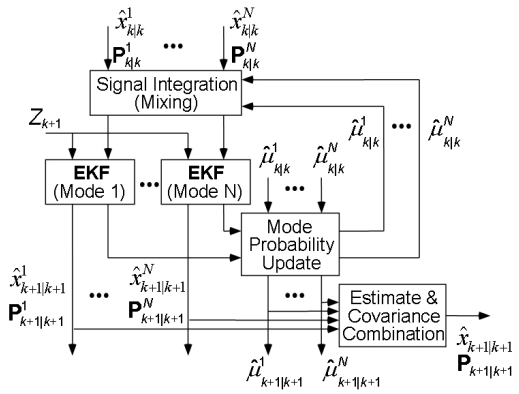

Figure 8. The data flow of the IMM algorithm.

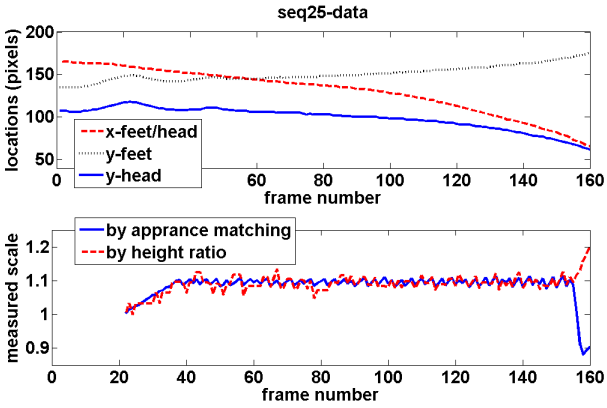

Figure 9. The original observation data from sequence number 25. Top: measured head and feet locations in the image; Bottom: measured scales by appearance matching and height ratio methods.

4.4. IMM Implementation

The multiple EKFs which correspond to multiple pedestrian height hypothesis are integrated under the IMM framework (for implementtaion details, see chapter 3 in [2]). Assuming there are N modes inside the IMM (Fig.8), the hypothesis height of the ith mode is H_i, and the observed head and feet locations in the image are $(x_{head}, y_{head}, x_{feet}, y_{feet})$, the initial state vector, $\mathbf{x}_{0|0}^{(i)}$ of that mode is given by

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

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. The initial probability of each mode is set as uniformly distributed, i.e. $1/N$.

Once the N filters are initialized, the interacting step mixes the probabilities and states of all the modes based on the mode transition parameters. By using the mixed results as inputs, the i^{th} mode EKF is updated by the observation in the current frame, and the corresponding likelihood value can be calculated by (7). After normalizing the updated likelihood values, the final state at current time can be estimated by combining the single mode estimates with different probabilities.

5. Results

We collected seven sequences with known pedestrian heights (1.66m, 1.70m, 1.86m and 1.92m) using a FIR sensor mounted on the bumper of a vehicle. 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 $Q(3,1)$=$Q(1,3) = Q(4,2) = Q(2,4) = 1.3889 \times 10^{-5}$, 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.9 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 3. 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. 9). Hence, the scales from height ratio are used by the filter whenever the confidence value from the appearance matcher is low.

Fig.10 plots the estimated state vector from the IMM fil-

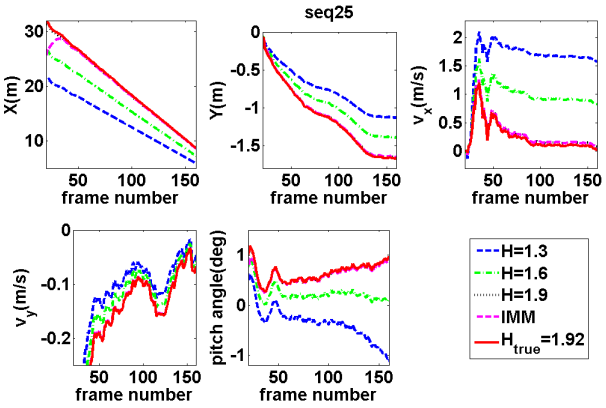

Figure 10. A comparison of the elements of the state vector from single mode EKFs and the IMM filter, where three of the single mode filters use the heights used in the IMM and one uses the ground truth height. This figure is best viewed in color.

ter 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. This was seen to be the case for all sequences in our test dataset which has pedestrians with several different ground-truth heights.

6. Conclusions

In summary, we have presented a novel application of the hierarchical model-based motion estimation framework to do temporal data-association, matching and scale estimation from detected pedestrian ROIs in FIR image sequences. The estimated scale allows us to use the multi-hypothesis-mode filtering algorithm to more accurately estimate the location of a pedestrian. Experiments demonstrate that this allows location estimates very close to those obtained from a known pedestrian height model.

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. Future work will focus on quantitatively evaluating the effect of these factors.

7. 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] 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*, pages 13–18, 2007.

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

[3] 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*, pages 194–199, 1994.

[4] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, volume 588, pages 237–252. Springer, 1992.

[5] 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*, pages 584–589, 2004.

[6] 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*, pages 462–467, 2006.

[7] P. Burt and E. Adelson. The Laplacian pyramid as a compact image code. *Communications, IEEE Transactions on [legacy, pre-1988]*, 31(4):532–540, 1983.

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

[9] D. Gavrila, J. Giebel, M. Perception, D. Res, and G. Ulm. Shape-based pedestrian detection and tracking. In *Intelligent Vehicle Symposium, 2002. IEEE*, volume 1, 2002.

[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*, pages 73–78, 2008.

[11] 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.

[12] 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*, pages 339–345, 2007.

[13] 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*, pages 37–42, 2008.

[14] R. Miezianko. Terravic research infrared database. IEEE OTCBVS WS Series Bench.

[15] 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*, 9(2):333– 342, 2008.

[16] 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)*, pages 28–39, 2004.

[17] 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.

[18] F. Xu, X. Liu, and K. Fujimura. Pedestrian detection and tracking with night vision. *IEEE Trans. on Intelligent Transportation Systems*, 6(1):63– 71, 2005.