Appearance and motion based data association for pedestrian tracking

Zhengqiang Jiang*, Du Q. Huynh*, William Moran†, and Subhash Challa†
*School of Computer Science and Software Engineering
The University of Western Australia, Perth, Australia
Email: {jiang, du}@csse.uwa.edu.au
†Department of Electrical and Electronic Engineering
The University of Melbourne, Melbourne, Australia
Email: wmoran@unimelb.edu.au, challas@ee.unimelb.edu.au

***Abstract*—In this paper, we present a method that combines appearance and motion information for data association in an interacting multiple model framework for tracking pedestrians in video sequences captured by a fixed camera. We formulate data association as a bipartite graph problem and employ the Munkres' algorithm to associate the new observations with the existing tracks. Two strategies for data association are designed and compared. The first strategy is to impose a threshold value on the total likelihood from the appearance and motion models as the edge weights of the bipartite graph. The second strategy is to use the combined motion model as a validation gate and the likelihood from the appearance model only as the edge weights. For the interacting multiple model tracking framework, we incorporate three motion models: a stationary model, a constant velocity model and a constant acceleration model. Our experimental results show that the second strategy gives better tracking results.**

I. Introduction

Recently, there has been a significant interest in the computer vision community on the tracking of pedestrians in videos. One of the common problems that visual tracking algorithms are exposed to is that pedestrian trajectories are likely to be lost because of missing or incomplete observations of pedestrians. An additional problem is that pedestrians are incorrectly identified due to wrong data association.

Numerous visual tracking techniques have been proposed in the literature to address these problems. A common problem shared by these tracking techniques is how to strengthen the data association stage so that the observations are correctly assigned to the tracks. Data association techniques, such as the nearest neighbour algorithm [1], the Joint Probability Data Association Filtering (JPDAF) [2] and the Multiple Hypothesis Tracking (MHT) [3], are among the commonly used methods. The nearest neighbour algorithm is designed to assign each track to the closest measurement using a distance measure. Although such an algorithm is computationally efficient, it is too simple to deal with complex situations where multiple pedestrians are present. The JPDAF technique associates all possible measurements to each track and weighs each innovation using the probability derived from the track for each measurement. Since the JPDAF technique assumes that the number of targets to be tracked is fixed, it cannot handle situations where new targets enter the scene or existing targets leave the scene. The MHT algorithm has the ability to track a variable number of targets at different video frames. The algorithm iteratively generates new hypotheses by comparing the predictions of old hypotheses and the actual measurements using a distance measure. Each hypothesis at the current frame consists of a set of tracks corresponding to the current measurements. Pruning techniques in each iteration are used to remove unlikely assignments. The main drawback of the algorithm is that the computational cost increases exponentially with the length of the tracking period. The hierarchical association [4] is another technique for object tracking. Such a technique classifies tracking problems into different levels and deals with simple level problems first.

In graph-based data association techniques, the tracked targets in the previous frame and the new observations in the current frame are represented as vertices; the likelihood values or similarity measures between these two entities are represented as weights of the edges. Chen et al. [5] use a graph matching algorithm to track multiple targets captured by a static camera. They employ a bipartite graph for their data association problem. The similarity between the targets and candidates are computed as the Kullback-Leibler distance between the corresponding colour histograms. Their method, however, cannot handle the situation when the two targets have similar appearance in the same video frame. Taj et al. [6] use a graph-based method to perform data association for tracking multiple objects in video sequences. Similarity measurements for data association are computed based on multi-features, such as the centroids, velocities, sizes and histograms of the tracked objects. More recently, Reilly et al, [7] present a bipartite graph for maximum matching of objects between video frames. They divide a scene into grid cells and then use the Hungarian algorithm [8] to associate the tracks to the observations in each cell. The weight of each edge in the bipartite graph contains object spatial and orientation information.

Our research work reported in this paper differs from those reviewed above in some respects. Our method uses bipartite graph that combines the appearance and motion information of each pedestrian. We use the Munkres' algorithm [9] to assign the existing tracks with their most likely observations

at the current time. To avoid incorrect data association due to detection errors of the pedestrian detector, we investigate two strategies for discarding unlikely matches between the tracked targets in the previous frame and the observations identified in the current frame. The first strategy (abbreviated as TTL) is to use the total likelihood from the appearance and motion models while imposing a use-defined threshold to eliminate unlikely associations. The second strategy (abbreviated as VAL) is to use validation gate (e.g., see [10]) from the motion models to rule out those edges which violate the motion models. The appearance model that we use in our tracking technique is smoothed 4D colour histograms computed for the pedestrian windows. For the motion models, we incorporate a stationary model, a constant velocity model and a constant acceleration model into an interacting multiple model (IMM) framework to deal with complex motions of pedestrians in video sequences.

The paper is organized as follows. Section 2 presents the pedestrian appearance model. We describe our pedestrian tracking algorithm in an interacting multiple model framework and the two strategies for data association in Section 3. Experimental results are given in Section 4. Finally, conclusion and future research direction are outlined in Section 5.

II. Pedestrian appearance model

Our pedestrian tracking technique employs the human detector proposed by Dalal and Triggs [11]. Their method involves the computation of histogram of oriented gradient (HOG) descriptors from overlapping image windows and the training of a linear support vector machine (SVM) on the HOG descriptors. They apply the SVM classifier to all detection windows of an input image at multiple scales to classify detection windows into 'non-pedestrian' and 'pedestrian'. The HOG descriptors are known to contain redundant information because of the different window sizes (i.e. different scales) being used. So when the SVM classifier classifies a region as 'pedestrian', many pedestrian windows of different scales are often associated with the same pedestrian. This inevitably also includes extra undesirable background areas in the pedestrian windows. The removal of background areas in the pedestrian window is therefore necessary in order to get a more accurate human appearance model.

Our pedestrian appearance model is a smoothed 4D colour histogram described in our previous work [12]. For completeness of the paper, we briefly summarize the steps below. We firstly extract a pedestrian bounding ellipse from the pedestrian window returned by the HOG human detector. The bounding ellipse has the same centroid as the pedestrian window's. To eliminate the extra background areas mentioned above, major and minor diameters of the bounding ellipse are made to be 20% smaller than the height and width of the pedestrian window. We then halve each pedestrian bounding ellipse into the upper and lower body parts and compute the colour histogram in the $L^*a^*b^*$ colour space for each part. Concatenation of the two colour histograms produces a 4D colour histogram for each pedestrian bounding ellipse. The 4D colour histograms are smoothed using kernel density estimation [13].

III. Pedestrian tracking

A. Interacting multiple model tracking

Our method incorporates multiple motion models in an interacting multiple model [14] framework to track pedestrians in video sequences. The IMM technique is designed to operate multiple filters in parallel and combine their outputs to obtain more accurate state vectors. Such a technique recursively estimates the state vectors and their error covariance matrices of moving targets over each time instant.

We use the Kalman filter for each motion model of the IMM technique. Let $Z(t)$ denote the measurement vectors up to time t. Let $\mathbf{x}_k(t|t)$ and $P_k(t|t)$ be the state estimate and error covariance matrix of a target for the k^{th} motion model at time t, given the measurement observed up to time t. Let $\mu_{k|m}(t|t)$ be the model mixing probability from the k^{th} to the m^{th} motion models $(k, m = 1, \ldots, M)$ at time t, given the measurement $Z(t)$. Let $\mu_k(t)$ be the probability of k^{th} motion model. In the Markov chain transition matrix, the element ρ_{km} denotes the transition probability from the k^{th} to m^{th} motion models. This matrix is fixed in the tracking process. Let $\Lambda_k(t)$ be the likelihood function for the k^{th} motion model at time t. Let the state transition and the measurement matrix for this motion model be denoted by $T_k(t)$ and $H_k(t)$, respectively. The noise vectors $\mathbf{w}_k(t)$ and $\mathbf{v}_k(t)$ are assumed to follow the Gaussian distribution with zero mean and covariance matrices $Q_k(t)$ and $R_k(t)$.

The IMM method allows the system to handle complex motion of a moving target as it can model a dynamic system with multiple models switching from one to another. We model the human motions using 3 motion models:

- Motion model 1: stationary model;
- Motion model 2: constant velocity model;
- Motion model 3: constant acceleration model.

The state vector of each pedestrian for each model is represented by a 11-dimensional vector. For instance, the state vector $\mathbf{x}_k^i(t) = (x_k^i(t), y_k^i(t), \dot{x}_k^i(t), \dot{y}_k^i(t), \ddot{x}_k^i(t), \ddot{y}_k^i(t), w_k^i(t), h_k^i(t), \dot{w}_k^i(t), \dot{h}_k^i(t), s_k^i(t))^\top \in \mathbb{R}^{11}$, for $i = 1, \cdots, N$, represents the state of the i^{th} pedestrian bounding ellipse at time t for the k^{th} motion model, where N is the number of pedestrian windows returned by the HOG human detector. Here, $(x_k^i(t),\ y_k^i(t))$ denotes the coordinates of the centroid; $(\dot{x}_k^i(t),\ \dot{y}_k^i(t))$ and $(\ddot{x}_k^i(t),\ \ddot{y}_k^i(t))$ denote the velocity and acceleration of the ellipse, respectively; $w_k^i(t)$ and $h_k^i(t)$ are the minor and major diameters of the pedestrian ellipse; parameters $\dot{w}_k^i(t)$ and $\dot{h}_k^i(t)$ represent the rates of change of these two diameters; $s_k^i(t)$ is the scale of the HOG descriptor returned by the HOG human detector. The matrix $P_k^i(t|t)$ is the covariance matrix of the state vector of the i^{th} pedestrian at time t for the k^{th} motion model, given the measurement observed up to time t. For the stationary motion model, the velocity and acceleration terms are ignored.

The IMM algorithm consists of several main stages in each cycle. In the first stage, the mixing probability $\mu_{k|m}^i(t-1)$ is

computed using the matrix ρ^i_{km} and vector $\mu^i_k(t-1)$ at time $t-1$ as follows:

$$\mu^i_{k|m}(t-1) = \frac{1}{c^i_m}\rho^i_{km}\mu^i_k(t-1), \tag{1}$$

where $c^i_m = \sum_{k=1}^{3}\rho^i_{km}\mu^i_k(t-1)$.

Each model's probability $\mu^i_m(t)$ is then computed using its likelihood function $\Lambda^i_m(t)$ and its probability in the previous frame. The likelihood function, which is assumed to be Gaussian distributed, and $\mu^i_m(t)$ are given by:

$$\Lambda^i_m(t) = \mathcal{N}(Z^j_m(t) - H^i_m(t)\mathbf{x}^i_m(t|t-1) - v^i_m(t); \mathbf{0}, H^i_m(t)P^i_m(t|t-1)H^i_m(t)^\top + R_m(t)), \tag{2}$$

$$\mu^i_m(t) = \frac{1}{c}\Lambda^i_m(t)\sum_{k=1}^{3}\rho^i_{km}\mu^i_k(t-1), \tag{3}$$

where c is a normalization constant. In our tracking method, the measurement is a vector in $\mathbb{R}^5$ containing the centroid, the major and minor diameters, and the HOG feature descriptor scale of the pedestrian bounding ellipse.

In the final stage, the state vectors and their associated error covariances of the three motion models are combined using the following equations:

$$\mathbf{x}^i(t|t) = \sum_{m=1}^{3}\mathbf{x}^i_m(t|t)\mu^i_m(t), \tag{4}$$

$$P^i(t|t) = \sum_{m=1}^{3}\{P^i_m(t|t) + [\mathbf{x}^i_m(t|t) - \mathbf{x}^i(t|t)][\mathbf{x}^i_m(t|t) - \mathbf{x}^i(t|t)]^\top\}\mu^i_m(t). \tag{5}$$

B. Data association

Multiple pedestrian tracking requires a data association step in order to correctly assign the tracked pedestrians in the previous frame with the observations in the current frame.

1) The likelihood from the appearance and motion models: The appearance information (i.e., colour histogram) of a target is widely used to solve the data association problem. However, it is not sufficient for a human tracking method to determine data association according to the appearance model only, e.g. a tracking method may fail if two people wear similar clothing. To overcome this limitation, we formulate data association as a bipartite graph problem and we incorporate the information from the appearance and the motion models to get the edge weights of the graph.

A bipartite graph $G = \langle V, E\rangle$ is a special graph where vertices of the graph are divided into 2 partitions, V_1 and V_2, such that $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \emptyset$; if $e = (v_1, v_2) \in E$ then $v_1 \in V_1$ and $v_2 \in V_2$. This graph is particularly suited for the data association problem in tracking as clearly the tracked targets that have already been established from previous video frames represent one partition of the vertices of the graph while the new observations are in the other partition.

The appearance likelihood function $\Lambda^{i,j}_a(t|t-1)$ specifies how likely the i^{th} pedestrian at time $t-1$ corresponds to the j^{th} pedestrian bounding ellipse returned from the HOG human detector at time t for the appearance model. The appearance likelihood function, which is assumed to be Gaussian distributed, i.e.,:

$$\Lambda^{i,j}_a(t|t-1) = \mathcal{N}(d^{i,j}_a(t|t-1); 0.5; \sigma^2), \tag{6}$$

where $d^{i,j}_a(t|t-1)$ represents the Hellinger distance [15] between the i^{th} tracked pedestrian at time $t-1$ and the j^{th} detected pedestrian bounding ellipse at time t and the σ is the standard deviation of the Gaussian distribution. The Hellinger distance is computed based on the Bhattacharyya coefficient $\rho^{i,j}(t|t-1)$ [16] between two 4D smoothed colour histograms from the i^{th} pedestrian at time $t-1$ corresponds to the j^{th} pedestrian bounding ellipse at time t using the following equation [15]:

$$d^{i,j}_a(t|t-1) = 2(1 - \rho^{i,j}(t|t-1)). \tag{7}$$

The scale factor 2 in the above equation is not included in our implementation. We use the Hellinger distance to evaluate the similarity measures since a smaller Hellinger distance indicates that the two 4D smoothed colour histograms are similar. On the contrary, a smaller Bhattacharyya coefficient indicates that the two colour histograms are dissimilar, which is not natural to interpret.

Similarly, the motion likelihood function $\Lambda^{i,j}_v(t|t-1)$ describes the likelihood of the j^{th} pedestrian bounding ellipse returned from the HOG human detector at time t given the prediction of i^{th} pedestrian at time $t-1$ from our IMM tracking technique. The motion likelihood function, which is also assumed to be Gaussian distributed, is given by:

$$V^{i,j}_v(t|t-1) = Z^j(t) - H(t)\mathbf{x}^i(t|t-1), \tag{8}$$

$$S^i_v(t|t-1) = H(t)P^i(t|t-1)H^\top(t) + R(t), \tag{9}$$

$$\Lambda^{i,j}_v(t|t-1) = \gamma\mathcal{N}(V^{i,j}_v(t|t-1); \mathbf{0}; S^i_v(t|t-1)), \tag{10}$$

where $V^{i,j}_v(t|t-1)$ is the innovation between the prediction of the i^{th} pedestrian at time $t-1$ and the j^{th} pedestrian bounding ellipse at time t; $S^i_v(t|t-1)$ is the innovation covariance of the i^{th} pedestrian; vector $Z^j(t)$ defines the j^{th} measurement; $\mathbf{x}^i(t|t-1)$ and $P^i(t|t-1)$ are the predictions of both the state vector and its associated error covariance matrix at time $t-1$; γ is a normalization factor so that $\Lambda^{i,j}_v(t|t-1)$ is equal to 1 when $V^{i,j}_v(t|t-1) = \mathbf{0}$.

The total likelihood $\Lambda^{i,j}(t|t-1)$ of the appearance and motion models is computed as follows:

$$\Lambda^{i,j}(t|t-1) = \alpha\Lambda^{i,j}_a(t|t-1) + (1-\alpha)\Lambda^{i,j}_v(t|t-1), \tag{11}$$

where $\alpha \in [0,1]$ is the weighting factor of the appearance likelihood on the total likelihood.

2) Data association: We use the Munkres' algorithm to solve the bipartite graph matching. Unlike the Hungarian algorithm, the Munkres' algorithm can deal with the situations where the partition V_1 and V_2 have different numbers of vertices. In our method, the weights of the edges are likelihood values and data association is resolved by finding the maximum flow in the graph. As the Munkres' algorithm works on minimizing a cost matrix, we need to negate all the matrix

elements first. So minimising this new matrix is the same as maximizing the original matrix. However, the Munkres' algorithm may fail to associate a pedestrian with the correct observation if false positives and /or negatives are returned from the HOG human detector. To overcome this limitation, we investigate two strategies as follows:

- Strategy 1: Thresholding on the total likelihood (TTL);
- Strategy 2: Validation gate and appearance likelihood (VAL).

The first strategy (see Fig. 1 (a)) aims to remove the edges that link the tracked pedestrians and the observations if the combined likelihood values are smaller than an appropriate threshold. The TTL strategy uses the combined likelihood values from the appearance and motion models as the weights of the edges in the bipartite graph. For the VAL strategy as shown in Fig. 1 (b), we use the validation gate from the combined motion model for each track to avoid associating these observations if the motion model is violated. That is, if the association results in a drastic change of velocity, then the association would be discarded. The three-sigma rule [17] is applied to the motion likelihood function to remove those edges which violate the motion models. In brief, the three-sigma rule specifies that if the innovation of the predicted state of a track and an observation is less than 3 standard deviations from the mean of the Gaussian distribution, then the corresponding edge of the bipartite graph would remain. The VAL strategy uses only the likelihood from the appearance model as the weights of the edges in the graph.

From Fig. 1, we can see that using the two strategies, different bipartite graphs can result. This may lead to different tracking outputs. Even if the two bipartite graphs have the same topology, their edge weights are different. The TTL strategy also has the weighting factor α (see Eq. (11)) that influences its performance.

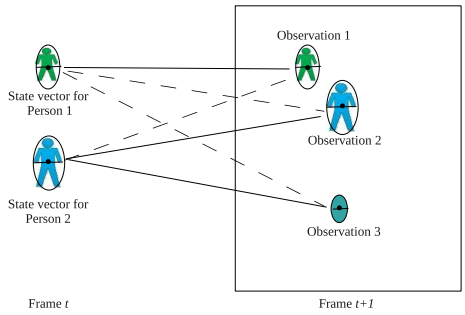
For the case where a tracked pedestrian is occluded and so no appropriate observation should be associated with this pedestrian, our IMM tracking technique would predict the pedestrian bounding ellipse using the combined motion model. Occlusion is dealt with using a sliding window of frames. The issue of occlusion is one of our research components still under investigation.

4. Experimental results

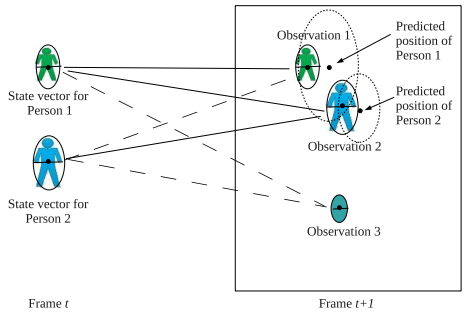
Our pedestrian tracking system is implemented in C++ together with the code available in the OpenCV library. Our tracking system has been tested on a 2.67 GHz processor PC.

We evaluate our tracking system on the two video databases obtained from the web. The first video set is selected from the CAVIAR video database [18], which has 384 $\times$ 288 pixels per frame. The second set is selected form the CVLAB video database [19], which has the 360 $\times$ 288 pixels per frame.

We initialize our IMM tracking framework using the pedestrian windows returned from the HOG human detector at the first video frame. Since the HOG human detector can return pedestrian windows with different centroid coordinates for a stationary pedestrian in consecutive frames due to noise,



(a) Thresholding on the total likelihood (TTL)



(b) Validation gate and appearance likelihood (VAL)

Fig. 1. Two strategies for removing edges before applying the Munkres' algorithm. The dashed lines shown here are edges that would be discarded if (a) using the TTL strategy their edge weights are below the pre-defined threshold; (b) using the VAL strategy the observations violate the combined motion models of the tracked persons. The validation gates are shown as dotted ellipses.

we assign the stationary motion model of our IMM framework with the prior probability set to 1 while setting the prior probabilities of both constant velocity and acceleration motion models to 0 at the first frame. The velocity and acceleration terms of the constant velocity and acceleration motion models are therefore 0. We assume that the initial state vectors and error covariance matrices for all three motion models are the same for our IMM framework. Assuming that the noise terms for all the components of the state vector are uncorrelated, the initial error covariance matrix of the k^{th} motion model for each pedestrian is defined to be $\hat{P}_k(0|0) = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_{\dot{x}}^2, \sigma_{\dot{y}}^2, \sigma_{\ddot{x}}^2, \sigma_{\ddot{y}}^2, \sigma_w^2, \sigma_h^2, \sigma_{\dot{w}}^2, \sigma_{\dot{h}}^2, \sigma_s^2)$. In all of our experiments, we empirically set $\sigma_x^2 = \sigma_y^2 = 0.8$, $\sigma_{\dot{x}}^2 = \sigma_{\dot{y}}^2 = 0.6$, $\sigma_{\ddot{x}}^2 = \sigma_{\ddot{y}}^2 = 0.3$, $\sigma_w^2 = \sigma_h^2 = 0.8$, $\sigma_{\dot{w}}^2 = \sigma_{\dot{h}}^2 = 0.6$ and $\sigma_s^2 = 0.7$. The switching probability matrix ρ_{km}^i from the k^{th} to m^{th} motion model is defined as:

$$\begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 0.05 & 0.8 & 0.15 \\ 0.05 & 0.15 & 0.8 \end{bmatrix}.$$

We set the switching probability between the constant velocity and acceleration models greater than the probability between the stationary model and the other two models since pedestrians are more likely to change the motion models between these two models.

We have conducted experiments of the IMM tracking method with the two different data association strategies described in the previous section. In the experiments reported

(a) Frame 1 (b) Frame 10 (c) Frame 44 (d) Frame 67 (e) Frame 84

Fig. 2. Tracking pedestrians in the presence of occlusion using the IMM tracking and the first strategy for data association.

(a) Frame 1 (b) Frame 10 (c) Frame 44 (d) Frame 67 (e) Frame 84

Fig. 3. Tracking pedestrians in the presence of occlusion using the IMM tracking and the second strategy for data association.

in this paper, we empirically define the likelihood threshold for the weight of each edge in the bipartite graph to be 0.7. Since our Hellinger distance values (see Eq.(7)) are in the range [0, 1], a truncated Gaussian distribution with mean at 0.5 having an area of 99.9% under the curve over this range requires the standard deviation to be around 0.15. We therefore set the standard deviation σ (see Eq. (6)) of the appearance likelihood function to this value. We chose α (see Eq.(11)) to be 0.8 to account for the uncertainty at times from the pedestrians' motions.

In Figs. 2 and 3, the performance of the TTL and VAL strategies are compared on the same video sequence. In frame 44 (Fig. 2 (c)), using the TTL strategy to construct the bipartite graph leads to an incorrect associations of person $\sharp 2$ and person $\sharp 3$ from the Munkres' algorithm. The reason for the wrong association is because of the large overlap of the pedestrian bounding ellipses and so the 4D colour histograms of both pedestrians are similar. As show in Fig. 4, this problem does not happen to the VAL strategy in frame 44 as the wrong associations are ruled out by the validation gate before Munkres' algorithm is applied. From these two figures we can see that tracking pedestrians using the VAL strategy outperforms the TTL strategy.

We also compared both strategies on the video sequences shown in Figs. 4 and 5. Using the same parameter values for σ and α as before, the tracking results are the same from both strategies. So we focus on describing the VAL strategy only from here on. As shown in Fig. 4, our method can successfully track pedestrians in the presence of occlusion. The HOG human detector returns one pedestrian window when person $\sharp 1$ occludes person $\sharp 2$ in frame 87. The bipartite graph generated by the VAL strategy results in correct assignment in frame 86. We then extract the coordinate of the centroid, the minor and major diameters and the scale of the HOG human detector from the pedestrian window to update the state estimates of these two pedestrians in frame 87. From this figure, we can see that the proposed tracking method can deal with cases such as when two people merge into a group and then split from the group.

We have verified our IMM tracking method using the ground truth and the camera calibration data provided in the CVLAB database [19]. The camera calibration data is used to map the ground truth positions of the people on the ground plane to the images using the camera calibration technique presented by Tsai [20]. The pedestrian trajectories generated from our tracking method and from the ground truth data are shown in Fig. 4. The closer are the pedestrian trajectories to the ground truth trajectories, the more accurate are the tracking results.

Figure 5 demonstrates our IMM human tracking using the VAL strategy for data association to handle false positive and negative errors from the HOG human detector. The first row of Fig. 5 shows pedestrian detection results from the HOG human detector. A false positive window can be seen in Fig. 5(c). On the other hand, the pedestrian window of person $\sharp 3$ is not successfully detected by the HOG human detector (see Fig. 5(d)). The false positive error is removed by data association using the VAL strategy and the false negative error is overcome by the prediction of the state vector for the missing person $\sharp 3$ from our IMM tracking framework. These two types of errors are successfully handled by our IMM tracking and the VAL strategy, as shown in Fig. 5(e)-5(h).

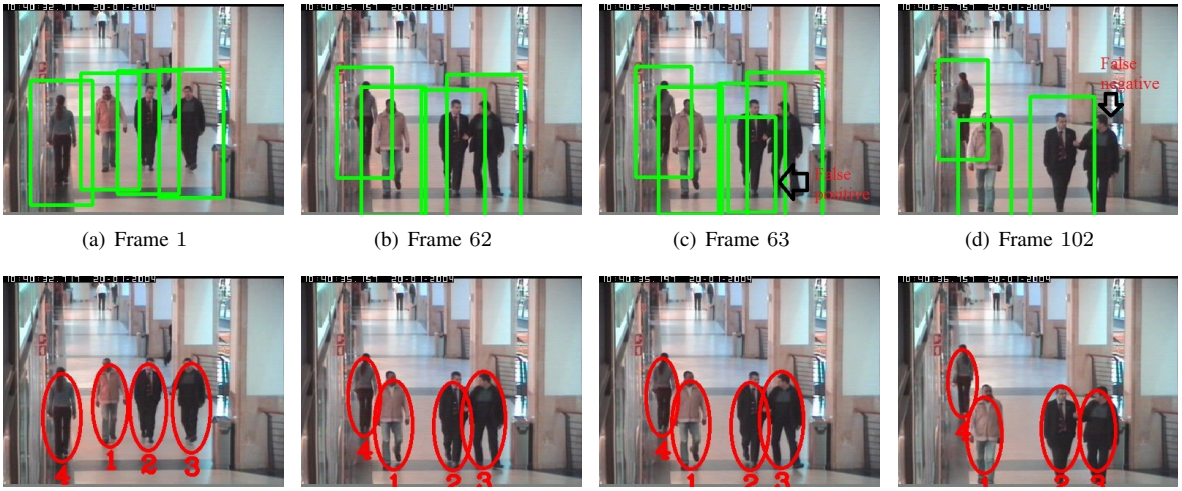
IV. Conclusion

We have presented a pedestrian tracking method that integrates an effective data association technique in an interacting multiple model framework for video sequences captured by a fixed camera. Our tracking method can cope with minor false positive and false negative errors from the pedestrian detector. Our method is also capable of tracking pedestrians under occlusion.

In our future work, we will look at video sequences with more crowded and complex scenes. Our human tracking method uses the HOG human detector that is trained using the pedestrian full body. The detector is therefore not suitable for detecting pedestrians that are too small in size (i.e., too far away from the camera). We intend to investigate pedestrian detectors that can deal with these cases and then use such detectors to track pedestrians in cluttered scenes.

(a) Frame 31 (b) Frame 87 (c) Frame 96 (d) Frame 125

Fig. 4. Tracking pedestrians in the presence of partial occlusion using the IMM tracking with the VAL strategy. The red curve represents the trajectory of person ♯1 while the blue curve is the image points projected from the ground truth data of person ♯1 on the ground plane. The magenta curve represents the trajectory of person ♯2 while the cyan curve is the image points projected from the ground truth data of person ♯2 onto the ground plane.



(a) Frame 1 (b) Frame 62 (c) Frame 63 (d) Frame 102

(e) Frame 1 (f) Frame 62 (g) Frame 63 (h) Frame 102

Fig. 5. Examples showing the handling of false positive and false negative errors by our IMM tracking and the VAL strategy for a video sequence containing four pedestrians. Row 1 shows the pedestrian detection results as green rectangles from the HOG human detector. A false positive error is present in (c) and a false negative error is present in (d). Row 2 shows the correct tracking results as red ellipses from our tracking method in the presence of these two types of errors.

ACKNOWLEDGMENT

Jiang would like to acknowledge the financial support of a PhD scholarship from the ARC Linkage Project grant LP0883417 and Sensen Networks Pty Ltd.

REFERENCES

[1] J. L. Crowley, P. Stelmaszyk, and C. Discours, "Measuring image flow by tracking edge-lines," in *International Conference on Computer Vision*, 1988.

[2] T. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *Oceanic Engineering, IEEE Journal of*, vol. 8, no. 3, pp. 173 – 184, Jul 1983.

[3] D. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 843 – 854, Dec 1979.

[4] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," *In European Conference on Computer Vision*, 2008.

[5] H. T. Chen, H. H. Lin, and T. L. Liu, "Multi-object tracking using dynamical graph matching," *Computer Vision and Pattern Recognition*, vol. 2, pp. II–210 – II–217, 2001.

[6] M. Taj, E. Maggio, and A. Cavallaro, "Multi-feature graph-based object tracking," *Multimodal Technologies for Perception of Humans*, pp. 190–199, 2007.

[7] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," *11th European Conference on Computer Vision*, 2010.

[8] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, pp. 83–97, 1955.

[9] J. Munkres, "Algorithms for assignment and transportation problems," *Journal of the Society for Industrial and Applied Mathematics*, vol. 5, no. 1, Mar 1957.

[10] A. Bissacco and S. Ghiasi, "Fast visual feature selection and tracking in a hybrid reconfigurable architecture," *Proceeding of the Workshop on Applications of Computer Vision*, 2006.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, Jun 2005.

[12] Z. Jiang, D. Q. Huynh, M. Moran, and S. Challa, "Tracking pedestrians using smoothed colour histograms in an interacting multiple model framework," *International Conference on Image Processing*, Sep 2011.

[13] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[14] H. A. P. Blom, "An efficient filter for abruptly changing systems," *The IEEE Conference on Decision and Control*, vol. 23, pp. 656 – 658, Dec 1984.

[15] G. Chaudri, "Bhattacharyya distance," *http://eom.springer.de/B/b110490.htm*, Mar 2009.

[16] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 142 – 149, 2000.

[17] M. S. Nikulin, "Three-sigma rule," *http://eom.springer.de/t/t092750.htm*, 2001.

[18] CAVIAR datasets, *http://homepages.inf.ed.ac.uk/rbf/CAVIAR/*.

[19] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267 – 282, Feb. 2008.

[20] R. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *Robotics and Automation, IEEE Journal of*, vol. 3, no. 4, pp. 323 – 344, Aug 1987.