Pedestrian Tracking by Associating Tracklets using Detection Residuals

Vivek Kumar Singh, Bo Wu, Ramakant Nevatia
University of Southern California, Los Angeles
{viveksin|bowu|nevatia}@usc.edu

Abstract

Due to increased interest in visual surveillance, various multiple object tracking methods have been recently proposed and applied to pedestrian tracking. However in presence of intensive inter-object occlusion and sensor gaps, most of these methods result in tracking failures. We present a two-stage multi-object tracking approach to robustly track pedestrians in such complex scenarios. We first generate high confidence partial track segments (tracklets) using a robust pedestrian detector and then associate the tracklets in a global optimization framework. Unlike the existing two-stage tracking methods, our method uses the unassociated low confidence detections (residuals) between the tracklets, which improves the tracking performance. We evaluate our method on the CAVIAR dataset and show that our method performs better than state-of-the-art methods.

1. Introduction

Tracking multiple objects is a well studied problem with an enormous literature and wide applications. One of the key applications that has recently gained a lot of attention is of tracking pedestrians in surveillance videos. [15], [14] present algorithms for multiple person tracking that seem to work reasonably well in many real life videos. However a common problem that these algorithms, and multiple object tracking algorithms in general, suffer from is that of track fragmentation due to missing observations, and faulty associations (identity switches) during inter object occlusions. Figure 1 shows some common problem cases in 2 track scenarios. To deal with these problems, researchers often use techniques such as delayed track association which involves associating observations to tracks over a temporal window (eg-MHT [11]). Since the number of potential observation-track associations increases exponentially with the window size, such methods are forced to use a bounded window size for higher efficiency. In practice, these methods robustly track objects across short sensor gaps and short occlusions, but fail as the gap/occlusion period gets close to or crosses the window boundary.

An alternative approach to deal with the above problems is to use a 2 stage tracking approach which generates *tracklets* (possibly partial target tracks) using standard tracking algorithms in first stage and then globally merges the generated tracks in a second stage. This method is referred to in literature as *track stitching/linking* [6], [10]. Recently such methods have been applied to track pedestrians and cars [6], [12], [7], [10], [1], [9]. These approaches, in general, generate tracklets using standard blob tracking methods and obtain the optimal association by maximizing the joint association likelihood of the tracklets. Based on the association framework, these approaches can be broadly categorized into graph-based and Bayesian inference-based methods.

Graph based methods approximate the joint likelihood of a set of tracklets by the product of pairwise tracklet association likelihoods of all pairs in the set. This reduces the problem of finding the optimal tracklet association to an assignment problem, which is then solved using the Hungarian algorithm. In [4], this method is used to associate tracks obtained from cameras with non-overlapping views. In [12], authors use this method to associate human target tracklets formed due to tracking failures within the scene; they simultaneously associate tracklets and estimate the source and sink regions (from where new track start and end respectively) in an EM framework. [10] extend this method to successfully track targets across multiple split and merge events or long sensor gaps; based on the tracklet association likelihoods they hypothesize merge and split events, and perform an iterative search over the merge/split hypotheses, accepting hypotheses that minimizes the joint tracklet

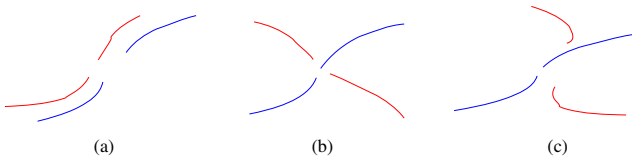

Figure 1. Common track interactions in multiple target tracking. (a) Sensor gap (no occlusion) (b) Inter object occlusion (c) Simultaneous gap and object occlusions. Segments of same color belong to the same object trajectory

likelihood at each step. [2] proposes another graph based approach which generates multiple hypotheses of object trajectories and greedily reject hypotheses based on the information from previous and later frames.

In [6], authors present Bayesian framework for linking tracklets (which they refer to as *strokes*). Each tracklet is represented as a node in a Bayesian network and has a state-observation model associated with it, where the state space represent the possible tracks that the tracklet may belong to. The nodes that may potentially belong to the same track are connected by an edge. The association problem is posed as a labeling problem (each label correspond to one target track) and the MAP association estimate is computed by max-marginalization. [5] extends this method to handle simultaneous merge and splits cases. [1] use event driven DBNs to guide the underneath tracklet association process; this is achieved by adding an event dependent term in the pairwise tracklet association likelihood model. [9] use a similar Bayesian network tracker to track soccer players over a long video sequence. In order to keep the inference computationally tractable, the approach requires explicit pruning of edges, for eg: dependencies between the tracklets that are distant in time are removed.

Above mentioned approaches have certain limitations, especially for surveillance scenarios. Firstly, the tracklets are generated using a motion blob tracker. Thus these approaches are sensitive to the background motion and a tracklet may correspond to a single object (when object blob is isolated) or a group of objects (when multiple object blobs merge possibly due to partial occlusion). Even though the association framework maintains the identity of objects across merge-split scenarios, the individual object tracks merged during tracking are not segmented. This is especially important in surveillance scenarios where the objects must be individually tracked to detect events such as loitering in a crowded scene. Secondly, the association likelihood model in above approaches does not take in account the low confidence observations in the gaps. These observations may be detections which failed to satisfy the track initiation criteria used by the tracker to generate the tracklets. We refer to these unassociated detections as *residuals*. In this work, we propose a 2 stage tracking method, *residual-based track linking*, which addresses the above mentioned issues and reliably tracks multiple objects in presence of sensor gaps and intensive inter object occlusions. We evaluate our system on the CAVIAR dataset [3] and show that our approach associate tracklets accurately and performs well even under complex scenarios.

1.1. Outline of our approach

Our approach differs from the existing approaches both in tracklet generation stage and the association likelihood model. Figure 2 shows a block diagram of our approach.

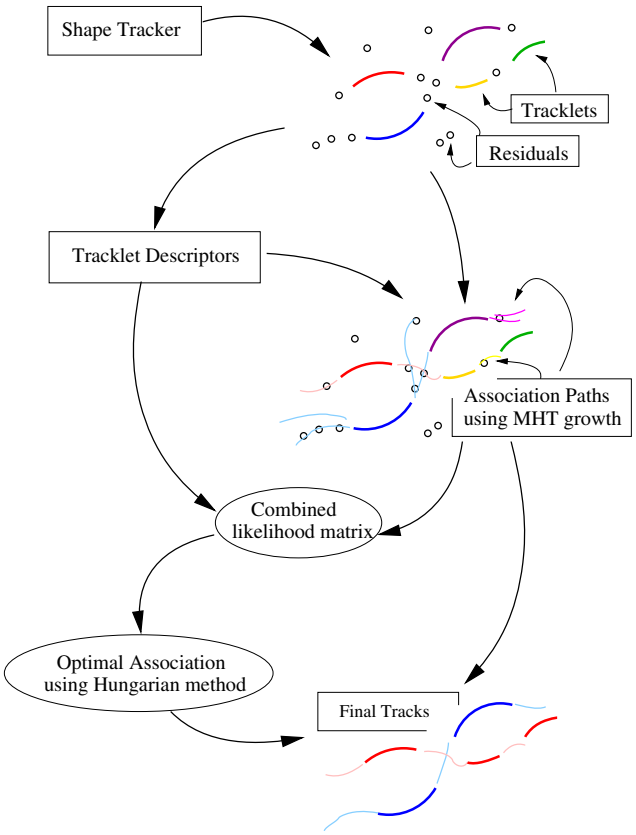

Figure 2. Block Diagram of our approach. The tracklets having same color belong to the same object trajectory. Association paths of a tracklet are shown by using the lighter shade of the color used for the tracklet (for eg - light blue segments are associations path generated using tracklet in deep blue).

Instead of using a blob tracker, we use a shape-detection based tracking approach to generate the tracklets; one tracklet is associated with only one object. Only high confidence detections are associated with tracklets while the low confidence detections are marked as residual observations and are used later in computing tracklet associations. From each tracklet we compute a *tracklet descriptor* consisting of the object color, motion model and height. Using this descriptor and residual observations, we *grow* each tracklet with a Multiple Hypothesis Tracker to find its potential associates. An *association path* is a chain of residual observations that connects 2 tracklets. We then compute the pairwise association likelihoods using tracklet descriptors and association paths separately, and combine them to get an overall association likelihood which in turn is used to compute the optimal association. The tracklet gaps between the associated tracks are then bridged using the potential association paths generated in the growth.

We could not find any existing literature comparing the two association approaches (Bayesian-inference based and

graph-based). Hence we conducted some preliminary experiments applying both methods on selected sequences from CAVIAR dataset [3]. We observed that both methods gave comparable association results, but the inference-based association method was much slower than graph-based association method. Hence in this work, we only describe the graph-based method.

Rest of the paper is structured as follows. Section 2 describes the tracklet generation approach. Section 3 describes the tracklet association framework. Section 4 discusses the evaluation results obtained on CAVIAR dataset, followed by the conclusion.

2. Tracklet Generation

We use a shape based pedestrian tracking method to generate the tracklets. In each frame, four body part detectors (described in [13]) are applied to detect pedestrians and their positive responses are combined to form human detection responses. We then track the pedestrians by associating the detection responses across consecutive frames. However if no match is found for a currently tracked object in a number of successive frames then the track is terminated. We initialize a new track whenever detection responses (with a confidence value above a certain threshold τ_d) can not be associated with currently tracked objects. All the part detection responses that do not get associated with a track are used as *residuals* during the tracklet association stage. Figure 3 illustrate the residuals and tracklets on a sample frame from the CAVIAR dataset [3].

A high value of the threshold τ_d results in fewer and possibly shorter tracklets as compared to tracklets generated using a lower value. However, a lower value of τ_d also produces more false alarm tracks. With fewer tracklets per object track, tracklet association method may become less effective for tracking; furthermore the chance of missing a track (number of objects with no tracklet associated to it) will also increase. Hence τ_d offers a trade-off between the reliability of tracklets and the effectiveness of using tracklet association for tracking.

(a) (b)

Figure 3. An example of tracklets and shape residuals (a) Sample frame from the CAVIAR dataset (b) tracklets and shape residual observations on that frame. The bounding boxes in red are residuals and those in yellow and magenta are tracklets.

3. Pairwise Tracklet Association

In this work, we adopt the tracklet association approach introduced in [12] and further studied in [7], [10]. This approach models the joint association likelihood $P(T_1, T_2 ... T_n)$ of tracklets as the product of pairwise association likelihoods $P(T_i, T_j)$. The pairwise association likelihoods are then represented in an association cost matrix $\mathcal{A}$ (with $A_{ij} = -log(P(T_i, T_j))$ and the optimal association (maximizing joint likelihood) is computed using the Hungarian algorithm.

For a set of n tracklets, association cost matrix $\mathcal{A}$ is an $(n+1) \times (n+1)$ matrix [1] defined as

$$\mathcal{A} = \begin{pmatrix} [\mathcal{A}_{1-1}(T_i, T_j)]_{n \times n} & [\mathcal{A}_{term}(T_i)]_{n \times 1} \\ [\mathcal{A}_{init}(T_j)]_{1 \times n} & 0 \end{pmatrix}$$

where, $\mathcal{A}_{1-1}(T_i, T_j) = -logP_{1-1}(T_i, T_j)$ is the pairwise association cost of tracklet T_i with T_j, $\mathcal{A}_{init}(T_j) = -logP_{init}(T_j)$ is the cost of initiating a track at tracklet T_j and $\mathcal{A}_{term}(T_i) = -logP_{term}(T_i)$ is the track termination cost at tracklet T_i.

We separately compute the association likelihood between every tracklet pair using the tracklet descriptors and the association paths, and combine them to get the overall association likelihood $P_{1-1}(T_i, T_j)$. Recall that *tracklet descriptors* consist of the object color, motion model and height, while *association paths* are chains of residual observations that connect 2 tracklets, and are obtained by *growing* tracklets during a second pass over the video. Note that the residual observations in an association path include both low confidence detections and hypotheses generated using the color model during the growth. We refer to the association likelihood computed using tracklet descriptors as $P^G(T_i, T_j)$ and using association paths as $P^L(T_i, T_j)$. Since we do not use the knowledge of potential entry and exit locations in the scene, the initiation and termination probabilities for a tracklet T_i are computed only using the association paths.

3.1. Tracklet Descriptors

The *tracklet descriptor* of a tracklet T_i is represented by the tuple $\{M_i, C_i, H_i\}$, where M_i is the motion model of the object, C_i is the color model of the object and H_i is the estimated 3D height of the object assuming the camera is stationary and calibrated, and object is in the known plane.

[1] This representation of association matrix $\mathcal{A}$ is common [12] and is used for notational convenience. To directly apply Hungarian algorithm to obtain the optimal association, $\mathcal{A}$ is defined as a $2n \times 2n$ matrix $\mathcal{A}'$ with $\mathcal{A}'_{1-1} = \mathcal{A}_{1-1}$; $\mathcal{A}_{init}$ and $\mathcal{A}_{term}$ as diagonal $n \times n$ matrices with $\mathcal{A}'_{init}(T_i, T_i) = \mathcal{A}_{init}(T_i)$ and $\mathcal{A}'_{term}(T_i, T_i) = \mathcal{A}_{term}(T_i)$.

Body Color Model C_i: We use a part based body color model $C_i = \{C_i^{ub}, C_i^{lb}\}$, where C_i^{ub} and C_i^{lb} are upper and lower body model respectively. Each part model is represented by the corresponding color histogram. To learn the color models, we segment the object using the foreground segmentation method [8] within the bounding box obtained from the shape detection response. We then use upper half and lower half of the detection response to update C^{ub} and C^{lb} respectively. Since the foreground segmentation is often not accurate, we associate a confidence value with each part model, say $conf^{ub}$ for C^{ub}, which indicates the accuracy of segmentation used to update the color model. For each detection response, we create an elliptical mask of the size of detection box. The confidence $conf^{ub}$ for part model C^{ub} is defined as $\frac{p}{p+3\times n}$, where p is the count of foreground pixels within the ellipse and n is the count of pixels outside. Note the extra weight to the negative overlapping pixels is added to penalize for including the pixels outside the ellipse.

Object Motion Model M_i: Like [10], we use a constant velocity model to predict the position the object from its tracklet. We use both forward and backward motion model for robustness. Motion model in forward direction is represented by a gaussian $\{x_i^F, \Sigma_i^F\}$ (mean-covariance pair), obtained using a Kalman filter over the last n detections of the tracklet. The motion model in backward direction $\{x_i^B, \Sigma_i^B\}$ is computed similarly.

Height Estimate H_i: We compute the a 3D height estimate of the object from the shape responses using the camera model. The expected 3D height averaged over the entire tracklet is used as the 3D height estimate. This height estimate is used to search for potential hypotheses during the tracklet growth (described in section 3.2).

Association Likelihood using Tracklet Descriptors

Following [10], we assume conditional independence of kinematic and color likelihood of tracklets and define pairwise association likelihood $P^G(T_i, T_j)$ as

$$P^G(T_i, T_j) = P_t^G(T_i, T_j) P_k^G(T_i, T_j) P_a^G(T_i, T_j)$$

where P_t^G is a temporal constraint function and P_k^G and P_a^G are kinematic and appearance based association likelihoods. $P_t^G(T_i, T_j)$ is a binary function which equals to 1 if tracklet T_i ends before T_j starts and 0 otherwise. $P_{init}^G(T_i)$ and $P_{term}^G(T_i)$ are both set to 0.

Kinematic [10] models kinematic likelihood of tracklet pair (T_i, T_j) as the product of position likelihoods of T_i where T_j starts and of T_j where T_i ends. We use a simple extension of their model by considering the kinematic likelihood for all frames between T_i and T_j and using the highest value as the kinematic likelihood of the tracklet pair. We define the kinematic likelihood between each tracklet pair as

$$P_k^G(T_i, T_j) = max_{t \in [e_i, s_j]} \quad \mathcal{N}(x_i^F(t), x_j^B(t), \Sigma_i^F(t)) \times \mathcal{N}(x_j^B(t), x_j^F(t), \Sigma_j^B(t))$$

where, $\mathcal{N}(x, \mu, \Sigma)$ is a Gaussian distribution with mean μ and variance Σ; $x_i^F(t)$ and $x_i^B(t)$ are predicted positions of object corresponding to tracklet T_i at frame t using forward and backward motion models respectively; s_j is the frame at which T_j initiates and e_i at which T_i terminates.

Appearance We define the color based association likelihood using the Gibbs model (similar to [7]). The weighted Bhattacharya distance between the tracklet color models is used as the distance measure $d(C_i, C_j)$ between the tracklets.

$$P_a^G(T_i, T_j) \quad = \quad e^{-k \times d(C_i, C_j)}$$

where,

$$d(C_i, C_j) \quad = \quad conf_i^{lb} \times conf_j^{lb} \times d(C_i^{lb}, C_j^{lb}) + conf_i^{ub} \times conf_j^{ub} \times d(C_i^{ub}, C_j^{ub})$$

3.2. Association Paths using Tracklet Growth

To compute the association likelihood using the residual observations, we find association paths between each tracklet and its potential associates. To achieve this, we initiate an object tracker at each tracklet. We refer to this as *growing a tracklet*. For each tracklet T_i, in order to simultaneously find all the association paths between T_i and its potential associates, a multiple hypothesis (MHT) tracker is used. MHT is a standard tracking technique which maintains multiple hypotheses for each target object and eventually select the best path as the object track. To facilitate tracklet associations during the growth, we increase the likelihood of hypothesis associations toward potential tracklet associates. In order to avoid hypotheses space explosion, we restrict the maximum number of hypothesis for each tracklet in a frame to $N_{max} = 4$ and also length of the tracklet to $W_{max} = 100$ frames (4 secs).

3.2.1 Tracklet Growth: Multiple Hypothesis Tracking

During the growth phase, the ith tracklet at frame f, denoted as T_i^f, is characterized by its hypotheses in that frame $\{h_{i,j}^f\}_{j=1}^{N_i}$, where N_i is the number of hypotheses of T_i^f. Each hypothesis $h_{i,j}^f$ constitutes the object motion model $M_{i,j}^f$, the color model C_i, a 3D height estimate of the object H_i and an exploration factor. The exploration factor determines the spatial transition threshold allowed between 2 detections in successive frames. Note that during the growth stage, we do not update the hypothesis color model C_i or the 3D height estimate H_i. During the data association stage of

growth, we ensure that each hypothesis is associated with only one hypothesis in previous frame, hence the growth results in a tree data structure.

Initialization: MHT growth for tracklet T_i is initiated at the last detection of the tracklet. We refer to this hypothesis as the *root* of the growth tree. The root hypothesis is initialized using the global statistics of T_i, linear motion model M_i and color model C_i.

Data Association(frame f to frame $f' = (f+1)$): For frame f', the color likelihood map using the color model and foreground segmentation from the foreground model are obtained. For jth hypothesis $h^f_{i,j}$ of each tracklet T_i, a likelihood region r for child hypotheses is obtained using the motion model $M^f_{i,j}$.

New hypotheses are generated both using pedestrian detector (*shape hypothesis*) and using the joint foreground-appearance model. Each shape hypothesis $h^{f'}_{shp}$ which overlaps r above a certain threshold is considered a potential child hypothesis (NOTE: $\{h^{f'}_{shp}\}$ includes both residual shape observations and tracklet shape observations in frame f'). To generate the foreground and appearance based hypotheses, an elliptical object mask (created using H_i) is convolved with region r in joint appearance and foreground map. Hypotheses corresponding to the maximas in this convolved space are added as potential child hypotheses.

Now each potential child hypothesis $h'(= h^{f'})$ is assigned an appearance likelihood probability $p_{app}(h')$ equal to average appearance likelihood of the pixels within the hypothesis. Similarly $p_{fg}(h')$, foreground appearance likelihood, is computed from foreground map. $p_{shp}(h')$, hypothesis shape likelihood, equals to the confidence of detection (max = 1). For foreground-appearance hypotheses, $p_{shp}(h') = 0$. Final probability of a hypothesis h is $p_{h'}$ is given by

$$p(h') = \lambda_s p_{shp}(h') + (1-\lambda_s)\frac{1}{2}(p_{fg}(h') + p_{app}(h'))$$

where, λ_s is relative weight given to the shape detections. A high value of λ_s encourages growth using mainly the shape hypotheses. Since we use all part detection responses as shape hypotheses, we get quite a few false alarms. So for our experiments, we set λ_s to a conservative value of 0.2.

The transition probability between a hypothesis $h(= h^f_i)$ and its child hypothesis $h'(= h^{f'}_i)$ is given by

$$p(h, h') = \mathcal{N}(h'_{pos}, h_{pos}, (\alpha h_{width})^2)$$

where, α is exploration factor of h. Low value of α restricts the spatial growth of hypothesis while a higher value encourages spatial exploration. In order to encourage branching (to find potential hypothesis), when a hypothesis in frame f branches in two 2 hypotheses in frame f', the one with lower probability is assigned a higher value of α.

Selecting Potential Associates A^p_i: Instead of using all the tracklets within a temporal window of the tracklet as potential associates, we determine potential associates for each tracklet T_i using the global statistics.

$$A^p_i = \{T_j | i \neq j, P^G(T_i, T_j) > \tau\}$$

To select the most probable associates, we use $\tau = 0.5 \times max_j P^G(T_i, T_j)$. In order to increase the association likelihood of a tracklet T_i and its potential associates A^p_i, during the growth of T_i the child shape hypotheses that corresponds to the tracklets in A^p_i, a higher value of α is used.

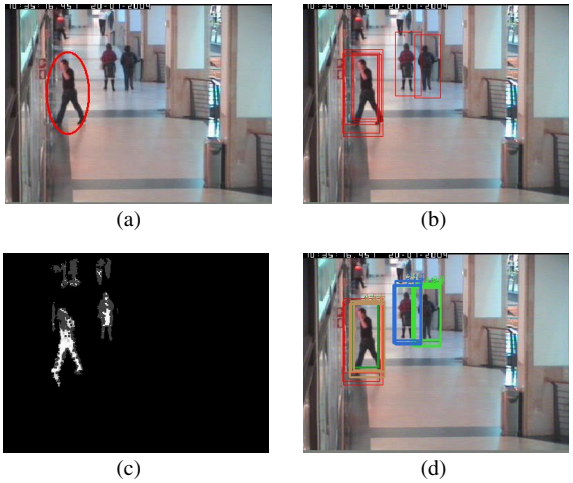

Figure 4. Multiple Hypothesis Growth (a) A sample image frame (b) Weak shape observations (residuals) in the frame (c) Combined foreground and appearance map for person encircled in red in figure(a). (d) Accepted hypotheses for growth

Pruning: Now each potential child hypothesis h' (with parent hypothesis h) is assigned a confidence value, $conf(h') = p(h, h') \times p_{h'}$. All the hypotheses that overlap (> 0.9) with another hypothesis of higher confidence value are removed. Then the hypothesis with least confidence is removed repeatedly until the number of child hypotheses becomes less or equal to N_{max}.

Once the growth is complete, the hypothesis tree is pruned to remove dangling branches (branches which do not associate with a tracklet and are less than 5 frames long).

3.2.2 Computing Association Path Likelihood

After tracklet growth is complete, the association likelihood between the tracklets is computed using the association paths formed during the growth.

Suppose growth tree of T_i associates with T_j, then we have an association path, $path(T_i, T_j)$ connecting tracklet T_i to T_j. The likelihood of the association path $P^L(T_i, T_j)$ is given by

$$P^L(T_i, T_j) = \overline{p_{shp}(h)} \times (\overline{p_{app}(h)})^{\frac{1}{2}} \times avg\_conf$$

where,

$$\overline{p_{shp}(h)} = \sum_{h \in path(T_i,T_j)} \beta \times p_{shp}(h)$$
$$\overline{p_{app}(h)} = \sum_{h \in path(T_i,T_j)} \beta \times p_{app}(h)$$
$$avg\_conf = \prod_{h \in path(T_i,T_j)} (conf(h))^{\beta}$$

where, $\beta = \frac{1}{|path(T_i,T_j)|}$ normalizes the confidence measures over the entire path length. This ensures that the shorter paths are not preferred over the longer ones.

A termination path of T_i is defined as a path in growth tree of T_i which does not associate with any other tracklet. A termination path is denoted as $path(T_i, \phi)$. Since there can be many such paths, we define the termination probability as $P^L_{term}(T_i) = max_{path(T_i,\phi)} P^L(T_i, \phi)$, and the path corresponding to maximum likelihood as the termination path.

The initiation probability $P^L_{init}(T_i)$ and initiation path are computed similarly from a tree obtained by the backward growth (reverse in time) of the tracklets.

3.3. Overall Association Likelihood

We combine the association likelihoods obtained using the tracklet descriptors and association paths to get the overall association likelihood. The pairwise association probability P_{1-1} is defined as

$$P_{1-1}(T_i, T_j) = \lambda P^G(T_i, T_j) + (1 - \lambda) P^L(T_i, T_j)$$

where, λ is a model parameter indicating relative strength of tracklet descriptor based association wrt the association paths. When λ is 0, only association paths are used to compute the optimal association, while when λ is 1, only tracklet descriptors are used. The overall initiation and termination probabilities $P_{init}(T_i), P_{term}(T_i)$ are defined similar to P_{1-1}. Since $P^G_{init}(T_i)$ and $P^G_{term}(T_i)$ are both 0, using only P^G ($\lambda = 1$) to find the optimal association results in faulty associations due to lack of track termination/initiation. On the other hand, if we only use the association paths to compute the optimal association ($\lambda = 0$) then tracklets which belong to the same object track but has a large sensor gap does not get associated. To encourage association, we set a high value of λ ($= 0.8$).

4. Results and Evaluation

We validated our approach on the CAVIAR dataset [3]. This dataset is captured from a static camera mounted on top of a corridor with camera pointing towards the corridor. Objects to be tracked are not occluded by the background (except at entry and exit points) but inter-object occlusion is intensive. The entire dataset contains 235 object tracks over 26 video sequences ($36,292$ frames) of frame size 384×288. Figure 5 shows some frames from the dataset.

To quantify the performance of our tracking system wrt existing systems, we compare our results with Wu and Nevatia [14] which is the only method that reports results on the entire CAVIAR dataset. [14] uses detection based tracking to track the objects, similar to our tracklet generation method. If a detection is not found for a currently tracked object then a meanshift tracker is used to follow that object. This allows us to compare the performances of a local ([14]) and a global approach (our method).

4.1. Performance metrics

To quantitatively evaluate our system, we use the performance metrics described in [10]; we summarize them here.

To compute the performance measures, an optimal association is computed between the set of generated tracks $\mathcal{T}$ and the ground truth tracks G. In order to quantify an association between ground truths and generated tracks, a distance metric $D(\mathcal{T}_i, G_j)$ between track $\mathcal{T}_i$ and ground truth track G_j is defined as the average Euclidean distance between the feet position of the hypotheses. Cost of an association A is then defined as the sum of distances between the generated tracks and the ground truth tracks associated in A. The optimal association A^* is the association that minimizes this association cost.

Given an optimal association, performance metrics *track completeness factor* (TCF) and *track fragmentation* (TF) are defined as

$$TCF = \frac{\sum_i \sum_{\mathcal{T}_j \in A(G_i)} |O(\mathcal{T}_j, G_i)|}{\sum_i |G_i|}$$
$$TF = \frac{\sum_i |A(G_i)|}{|\{G_i | A(G_i) \neq \phi\}|}$$

where, $A(G_i) = \{\mathcal{T}_k\}$ is the set of generated tracks associated with G_i in the optimal association A^*, and $O(\mathcal{T}_j, G_i)$ is the set of frames where $\mathcal{T}_j$ and G_i overlaps, $|.|$ denotes the cardinality of the set. TCF indicates fraction of object tracks tracked correctly over the entire set, while TF indicates the average number of track fragments that constitute a track. A fragmentation score of 1 is ideal.

4.2. Results on CAVIAR dataset

Our shape-based pedestrian detector detects pedestrians with size greater than 20 pixel height. Hence in order to evaluate the tracking performance, we remove the ground truth tracks/detections where the size of pedestrian is too small. Figure 5(a) shows the rectangular mask used for filtering hypotheses. The filtering process removed 23 tracks out of the total of 235.

Our tracklet generation method generates 654 tracklets over the entire dataset, with 18 false alarm tracklets and 9 ID switches. Average size of tracklets is about 77 frames and about 20% of the tracklets are less than 10 frames long. The tracklets generated only associate with 172 ground truth tracks. Since our current method relies on the tracklets for track initiation, we are unable to track the remaining 50 tracks. During the association stage, 12 out of 18 false alarm tracks did not associate with any other track while the rest 6 did associate and resulted in track fragmentation.

Table 1 shows the comparative results. Method A correspond to the generated tracklets. Method B is a hypothetical approach where tracklets are associated using the ground truth tracks and then linearly interpolated to fill the tracklet gaps. The low TCF score of method B clearly indicate that approximating the tracklet gaps with linear trajectories is not enough and justifies the idea of using growth to fill the gaps between the tracklets associated with the same track. Method C in table 1 corresponds to Wu and Nevatia's method [14]. Method D correspond to association obtained using only the tracklet descriptors ($\lambda = 1$ in our method) while Methods E and F correspond to our method with λ set to 0 and 0.8 respectively. Recall that λ is the relative weight given to the association likelihood obtained using association paths wrt that obtained using tracklet descriptors. As expected the use of low value of λ results in increased fragmentation. Observe that Methods E and F both gives a much better tracking performance than Method C (with a higher TCF and a lower TF score). Since method F uses a global association framework, a lower TF score as compared to method C is expected. But notice that by using association paths to fill the tracklet gaps, both method E and F gives a much better TCF score than C and D.

Figure 5 shows some results generated by our method on the CAVIAR dataset. To differentiate the residual detections with the tracklet detections, the tracklet detection boxes are marked with a red dot at the bottom. Figure shows the qualitative accuracy of the residual observations generated during the growth.

	Number of tracks	TCF	TF
A. Tracklets	172	0.441	3.76
B. GT + Linear Interp	172	0.512	1.0
C. Wu & Nevatia [14]	185	0.667	1.2
D. Tracklet Descriptors	172	0.376	1.25
E. Ours ($\lambda = 0.0$)	166	0.726	1.41
F. Ours ($\lambda = 0.8$)	166	0.714	1.15

Table 1. Performance measures on CAVIAR dataset [3].

5. Conclusion and Future Work

In this work, we propose a *residual-based tracklet linking* algorithm to track objects in presence of intensive inter-object occlusions and sensor gaps. We use a shape based tracker to generate tracklets in the first pass and then associate the tracklets using tracklet descriptors and potential association paths between the tracklets. To validate our method, we apply it on the CAVIAR dataset and show that our method performs better than [14].

One limitation of our current system is that we use tracklets for track initiation and hence any track for which no tracklet is generated cannot be tracked by our method. Another limitation is that our method does not fix the identity switches that occur within the tracklets. We intend to address these issues in future. One way to deal with the initiation problem is to use tracklet growth to initiate a track using the residuals that did not get associated after tracklet linking is done.

Acknowledgements

This research was funded, in part, by the U.S. Government VACE program.

References

[1] M. Chan, A. Hoogs, R. Bhotika, A. Perera, J. Schmiederer, and G. Doretto. Joint recognition of complex events and track matching. In *Proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1615–1622, 2006. 1, 2

[2] A. Y. S. Chia, W. Huang, and L. Li;. Multiple objects tracking with multiple hypotheses graph representation. In *Proceedings of the IEEE international conference on Pattern Recognition (ICPR)*, volume 1, pages 638–641, 2006. 2

[3] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/. 2, 3, 6, 7, 8

[4] T. Huang and S. Russell. Object identification in bayesian context. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 1997. 1

[5] T. Huang and S. Russell. Tracking with bayesian networks: extension to arbitrary topologies. In *Proceedings of the International Conference on Image Processing (ICIP)*, pages II – 402–405, 2005. 2

[6] P. M. Jorge, J. S. Marques, and A. J. Abrantes. Online tracking groups of pedestrians with bayesian networks. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS-ECCV)*, pages 65–72, 2004. 1, 2

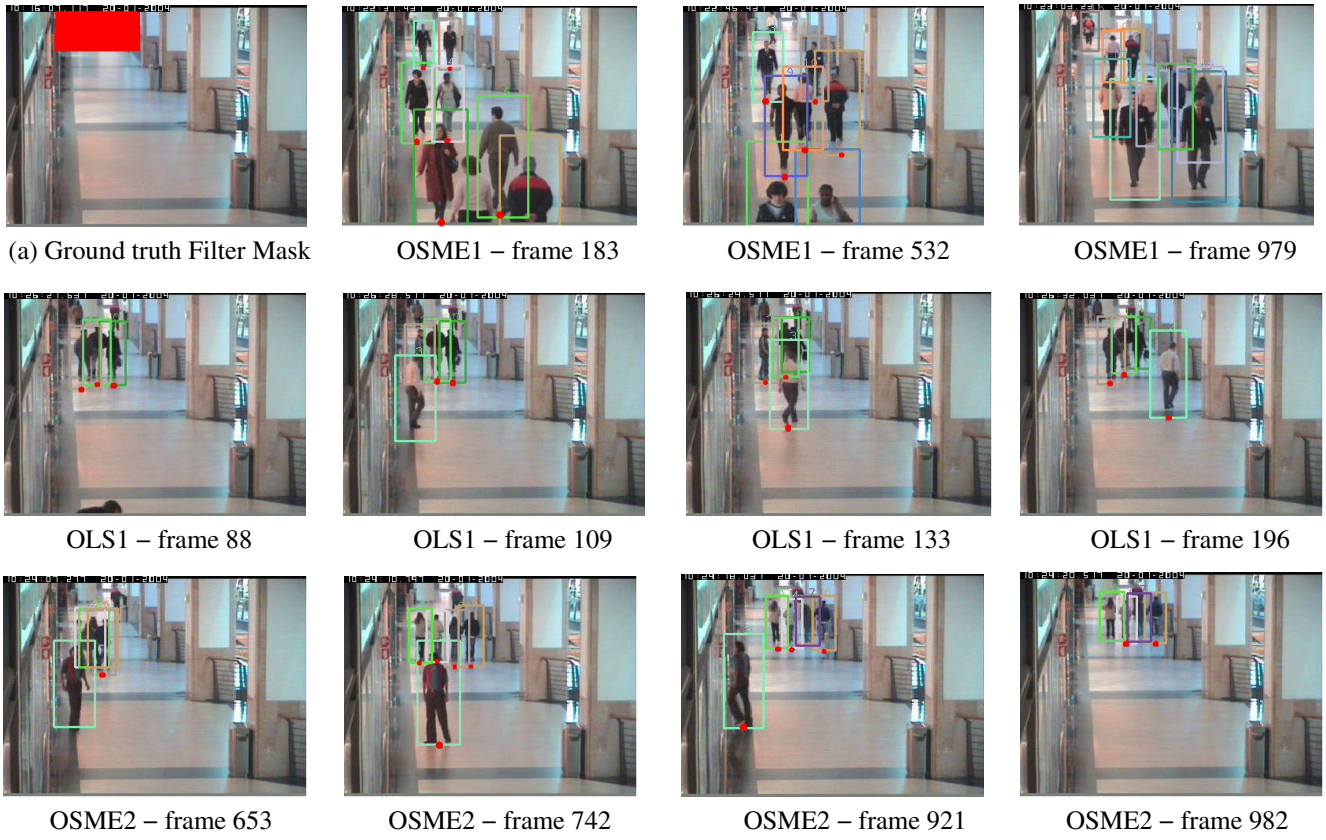

Figure 5. (a) shows CAVIAR background scene with the mask (in red) used to filter out the short ground truth object. Other images show sample results generated by our method on the CAVIAR dataset [3]. OSME1, OLS1 and OSME2 are OneStopMoveEnter1, OneLeaveShop1 and OneStopMoveEnter2 sequences from the CAVIAR dataset. Detection boxes with same color belong to the same object. Those that belong to a tracklet are marked with a red dot at the bottom of the box.

[7] R. Kaucic, A. Amitha Perera, G. Brooksby, J. Kaufhold, and A. Hoogs. A unified framework for tracking through occlusions and across sensor gaps. In *Proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 990–997, 2005. 1, 3, 4

[8] L. Li, W. Huang, I. Y. H. Gu, and Q. Tian. Foreground object detection from videos containing complex background. In *Proceedings of the eleventh ACM international conference on Multimedia*, 2003. 4

[9] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking - linking identities using bayesian network inference. In *CVPR '06: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 2187–2194, 2006. 1, 2

[10] A. Perera, C. Srinivas, A. Hoogs, G. Brooksby, and W. Hu;. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In *Proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 666–673, 2006. 1, 3, 4, 6

[11] D. Reid. An algorithm for tracking multiple targets. In *IEEE Trans. Automatic Control*, pages 24(6): 843–954, 1979. 1

[12] C. Stauffer. Estimating tracking sources and sinks. In *Proceedings of the IEEE Event Mining Workshop*, 2003. 1, 3

[13] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *ICCV*, pages 90–97, 2005. 3

[14] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *Proceedings of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 951–958, 2006. 1, 6, 7

[15] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR04*, pages II: 406–413, 2004. 1