Robust Object Tracking by Hierarchical Association of Detection Responses

Chang Huang, Bo Wu, and Ramakant Nevatia

University of Southern California, Los Angeles, CA 90089-0273, USA
{huangcha,bowu,nevatia}@usc.edu

Abstract. We present a detection-based three-level hierarchical association approach to robustly track multiple objects in crowded environments from a single camera. At the low level, reliable tracklets (i.e. short tracks for further analysis) are generated by linking detection responses based on conservative affinity constraints. At the middle level, these tracklets are further associated to form longer tracklets based on more complex affinity measures. The association is formulated as a MAP problem and solved by the Hungarian algorithm. At the high level, entries, exits and scene occluders are estimated using the already computed tracklets, which are used to refine the final trajectories. This approach is applied to the pedestrian class and evaluated on two challenging datasets. The experimental results show a great improvement in performance compared to previous methods.

1 Introduction

Tracking of objects is important for many computer vision applications. This is a relatively easy task when the objects are isolated and easily distinguished from the background. However, in complex and crowded environments, many objects are present that may have similar appearances, and occlude one another; also occlusions by other scene objects are common. We propose a method that can robustly track multiple objects under such challenging conditions.

Traditional feature-based tracking methods, such as those based on color [1], salient points [2] or motion blobs [3], do not have a discriminative model that distinguishes the object category of interest from others. Use of object detectors as discriminative models helps overcome this limitation. It also enables automatic initialization and termination of trajectories. However, though there has been significant improvement in object detection techniques, e.g. [4,5,6], the accuracy of the state-of-the-art object detectors is still far from perfect. Missed detections, false alarms and inaccurate responses are common (e.g. Fig.1). The tracking method must function with such failures, and also with the difficulties due to occlusions and appearance similarity among multiple objects.

In most existing association-based tracking methods, e.g. [6,7,8], an affinity score between detection responses (or tracklets) is computed once and fixed for all later processing. What we propose is a more flexible approach where associations are made in several levels and the affinity measure is refined at each level based on the knowledge obtained at the previous level. A scene model is also estimated from the tracklets and then used to generate the final object trajectories. This method is applied to track pedestrians in a number of videos; the experimental results show that our method achieves a large improvement in tracking accuracy without much increase in computational cost.

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part II, LNCS 5303, pp. 788–801, 2008.

Fig. 1. Misleading information provided by the pedestrian detector using the method of [6]

1.1 Related Work

The key issue of association based tracking is how to obtain correct associations robustly with noisy detection results. Multi-Hypothesis Tracking (MHT) [11], Joint Probabilistic Data Association Filters (JPDAFs) [12], and the particle filter based trackers [14,9] solve this problem by maintaining multiple hypotheses until enough evidence can be collected to resolve the ambiguity. Some other methods try to get better association by doing global analysis. For example, in [13], detection and tracking are coupled by Quadratic Boolean Programming that is solved by an EM-style algorithm. However, these approaches are limited for long-time association due to the combinatorial growth of hypothesis search space. Sampling methods such as MCMC [15] have also been employed to find the approximate solution in the high-dimensional hypothesis space.

The Hungarian algorithm [10] is another widely used approach for association. Wu and Nevatia [6] define an affinity measure based on position, size and color, and use the Hungarian algorithm to associate object hypotheses and detection responses at neighboring frames. Stauffer [7] first obtains tracklets by performing a conservative frame-to-frame correspondence, and then associates these tracklets by the Hungarian algorithm with an extended transition matrix that considers initialization and termination of each tracklet. He proposes a parametric source/sink model for the tracklet initialization and termination, and solves the coupled problem (estimating the source/sink model and associating tracklets) in an EM framework. Perera et al. [8] modify this extended transition matrix to maintain object identities even if their trajectories are merged or split. Kaucic et al. [17] extend Stauffer's method by introducing a segmentation based scene understanding module to estimate the locations of scene occluders. However, training the scene understanding module requires manually-labeling the background in a few frames of the video, which makes their approach not fully automatic. Singh et al.[16] use a Multiple Hypothesis Tracker to grow tracklets before associating them by the Hungarian algorithm. In these methods, the transition matrix for the Hungarian algorithm is computed only once and fixed for association. Hence, the errors in affinity computation caused by inaccurate detections are hard to be alleviated during the association process and are likely to be propagated to the higher level analysis.

1.2 Outline of our Approach

In our approach, object trajectories are obtained by progressively associating detection responses in a three-level hierarchical framework. As shown in Table.1, different models and methods are used to meet the requirements of association at different levels.

Table 1. Models and methods used in different levels of the hierarchical framework

	Motion	Appearance	Association	Scene model	Coordinates
Low level	N/A	Raw	Direct Link	N/A	Image
Middle level	Dynamic	Refined	Hungarian	General	Image
High level	Dynamic	Refined	Hungarian	Specific	Ground plane

First, at the low level, reliable tracklets are generated by linking detection responses in consecutive frames. A conservative two-threshold strategy is used to prevent "unsafe" associations until more evidence is collected to reduce the ambiguity at higher levels.

Second, at the middle level, the short tracklets obtained at the low level are iteratively associated into longer and longer tracklets. This is formulated as a MAP problem that considers not only initialization, termination and transition of tracklets but also hypotheses of tracklets being false alarms. In each round, positions and velocities of each input tracklet are estimated. This information helps refine the appearance model, and additionally provides a motion model to characterize the target. A modified transition matrix is computed and sent to the Hungarian algorithm to obtain optimal association.

Finally, at the high level, a scene structure model, including three maps for entries, exits and scene occluders, is estimated based on the tracklets provided by the middle level. Afterward, the long-range trajectory association is performed with the help of the scene knowledge based reasoning to reduce trajectory fragmentation and prevent possible identity switches.

In this framework, inaccuracies of detection responses, automatic initializations and terminations of trajectories, and occlusions by scene occluders are all taken into account. The main contributions of this paper include: 1) a novel three-level hierarchical framework to progressively associate detection responses, in which different methods and models are adopted to improve tracking robustness; 2) a modified transition matrix for the Hungarian algorithm to solve the association problem that considers not only initialization, termination and transition of tracklets but also false alarm hypotheses; 3) a novel Bayesian inference approach to automatically estimate a scene structure model as the high-level knowledge for the long-range trajectory association.

The rest of this paper is organized as follows: the proposed hierarchical association framework is elaborated in Section 2; experimental results on two challenging public video sets are shown in Section 3; some conclusions are made in Section 4.

2 Hierarchical Association of Detection Responses

We denote a detection response by $\mathbf{r}_i = (x_i, y_i, s_i, t_i, a_i)$, in which (x_i, y_i) is the position, s_i is the size, t_i is the occurrence frame index, and a_i is the color histogram; $T_k = \{\mathbf{r}_{k_i} | \forall i, t_{k_i} < t_{k_{i+1}}\}$ is an object trajectory/tracklet; $\mathcal{T} = \{T_k\}$ is the object trajectory/tracklet set; $\mathcal{T}^{\mathcal{L}}$, $\mathcal{T}^{\mathcal{M}}$ and $\mathcal{T}^{\mathcal{H}}$ are association results of the low level, the middle level and the high level respectively. We assume that a detection response can only belong to one tracklet/trajectory, which is formulized as a non-overlap constraint:

$$\forall \mathcal{E} \in \{\mathcal{L}, \mathcal{M}, \mathcal{H}\} \ \forall T_i, T_j \in \mathcal{T}^{\mathcal{E}}, \ T_i \cap T_j = \emptyset. \tag{1}$$

2.1 Low-Level Association

Denote by $\mathcal{R} = \{\mathbf{r}_i\}$ the set of all detection responses. The low-level association takes $\mathcal{R}$ as the input, and generates reliable tracklets by a simple and conservative method, *direct link*. Similar to that in [6], link probability between two responses is defined as the product of three affinities based on position, size and appearance:

$$P_{link}(\mathbf{r}_j|\mathbf{r}_i) = \begin{cases} A_{pos}(\mathbf{r}_j|\mathbf{r}_i)A_{size}(\mathbf{r}_j|\mathbf{r}_i)A_{appr}(\mathbf{r}_j|\mathbf{r}_i), & \text{if } t_j - t_i = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

Notice that the association can only happen between two consecutive frames. According to the non-overlap constraint in Equ.1, $(\mathbf{r}_i, \mathbf{r}_j)$ and $(\mathbf{r}_k, \mathbf{r}_l)$ are regarded as two *conflicting pairs* if $i = k$ or $j = l$. To prevent "unsafe" associations, two responses are linked if and only if their affinity is high enough and significantly higher than the affinity of any of their conflicting pairs:

$$\begin{aligned} &P_{link}(\mathbf{r}_i|\mathbf{r}_j) > \theta_1, \text{ and } \forall \mathbf{r}_k \in \mathcal{R} - \{\mathbf{r}_i, \mathbf{r}_j\}, \\ &\min\Big[P_{link}(\mathbf{r}_i|\mathbf{r}_j) - P_{link}(\mathbf{r}_k|\mathbf{r}_j), P_{link}(\mathbf{r}_i|\mathbf{r}_j) - P_{link}(\mathbf{r}_i|\mathbf{r}_k)\Big] > \theta_2, \end{aligned} \tag{3}$$

where θ_1 and θ_2 are the two thresholds.

Based on this *two-threshold strategy*, the low-level association can efficiently generate a set of reliable tracklets $\mathcal{T}^{\mathcal{L}} = \{T_k^L\}$. Any isolated detection response, which is not linked with any other one, is considered as a degenerate tracklet and also included in $\mathcal{T}^{\mathcal{L}}$. The low-level association does not resolve the ambiguity of conflicting pairs, as they can be figured out more effectively at higher levels.

2.2 Middle-Level Association

The middle level association is an iterative process: each round takes the tracklets generated in the previous round as the input and does further association. Take the first round, whose input is $\mathcal{T}^{\mathcal{L}}$, for example. We define a *tracklet association* as a set of tracklets: $S_k = \{T_{i_0}^L, T_{i_1}^L, \cdots, T_{i_{l_k}}^L\}$, where l_k is the number of tracklets in S_k. $T_k^M = \bigcup_{T_i^L \in S_k} T_i^L$ is the corresponding trajectory of S_k, and $\mathcal{S} = \{S_k\}$ is the *tracklet association set*.

The objective of the first round association can be formulated as a MAP problem:

$$\begin{aligned} \mathcal{S}^* &= \arg\max_{\mathcal{S}} P(\mathcal{S}|\mathcal{T}^{\mathcal{L}}) = \arg\max_{\mathcal{S}} P(\mathcal{T}^{\mathcal{L}}|\mathcal{S})P(\mathcal{S}) \\ &= \arg\max_{\mathcal{S}} \prod_{T_i^L \in \mathcal{T}^{\mathcal{L}}} P(T_i^L|\mathcal{S}) \prod_{S_k \in \mathcal{S}} P(S_k), \end{aligned} \tag{4}$$

assuming that the likelihoods of input tracklets are conditionally independent given $\mathcal{S}$, and the tracklet associations $\{S_k\}$ are independent of each other.

A Bernoulli distribution is used to model the probability of a detection response being a true detection or a false alarm. Let β be the precision of the detector, the likelihood of an input tracklet is defined as

$$P(T_i^L|\mathcal{S}) = \begin{cases} P_+(T_i^L) = \beta^{|T_i^L|}, & \text{if } \exists S_k \in \mathcal{S}, T_i^L \in S_k \\ P_-(T_i^L) = (1-\beta)^{|T_i^L|}, & \text{if } \forall S_k \in \mathcal{S}, T_i^L \notin S_k \end{cases} \tag{5}$$

where $|T_i^L|$ is the number of detection responses in T_i^L, and $P_+(T_i^L)$ and $P_-(T_i^L)$ are the likelihoods of T_i^L being a true detection and a false alarm respectively.

The tracklet association priors in Equ.4 are modeled as Markov Chains:

$$P(S_k) = P_{init}(T_{i_0}^L)P_{link}(T_{i_1}^L|T_{i_0}^L)\cdots P_{link}(T_{i_{l_k-1}}^L|T_{i_{l_k}}^L)P_{term}(T_{i_{l_k}}^L) \tag{6}$$

composed of an initialization term $P_{init}(T_{i_0}^L)$, a termination term $P_{term}(T_{i_{l_k}}^L)$ and a series of transition terms $P_{link}(T_{i_{i+1}}^L|T_{i_i}^L)$. Definitions of these terms will be given later in the subsection of implementation details.

Constrained by the non-overlap assumption in Equ.1, T_i^L cannot belong to more than one S_k. Thus, we rewrite Equ.4 by inserting $P_+(T_i^L)$ into its corresponding chain:

$$\begin{aligned}\mathcal{S}^* = \arg\max_{\mathcal{S}} \prod_{\forall S_k\in\mathcal{S}, T_i^L\notin S_k} P_-(T_i^L) \prod_{S_k\in\mathcal{S}} \Big[& P_{init}(T_{i_0}^L)P_+(T_{i_0}^L) \\ & P_{link}(T_{i_1}^L|T_{i_0}^L)\cdots P_{link}(T_{i_{l_k}}^L|T_{i_{l_k-1}}^L)P_+(T_{l_k}^L)P_{term}(T_{i_{l_k}}^L)\Big]. \end{aligned} \tag{7}$$

This MAP formulation has a distinct property compared to the previous work [7,8]: it allows $\mathcal{S}^*$ to exclude some input tracklets, rejecting them as false alarms instead of receiving their initialization/termination penalties or transition terms by linking them.

Hungarian Algorithm. Supposing there are n input tracklets, we can transfer the MAP problem in Equ.7 into a standard assignment problem by defining a transition matrix:

$$\mathbf{C} = \left[\begin{array}{cccc|cccc} C_{11} & C_{12} & \cdots & C_{1n} & C_{1(n+1)} & -\infty & \cdots & -\infty \\ C_{21} & C_{22} & \cdots & C_{2n} & -\infty & C_{2(n+2)} & \cdots & -\infty \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} & -\infty & -\infty & \cdots & C_{n(2n)} \\ \hline C_{(n+1)1} & -\infty & \cdots & -\infty & 0 & 0 & \cdots & 0 \\ -\infty & C_{(n+2)2} & \cdots & -\infty & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ -\infty & -\infty & \cdots & C_{(2n)n} & 0 & 0 & \cdots & 0 \end{array}\right]_{2n\times 2n} \tag{8}$$

whose components are defined as

$$C_{ij} = \begin{cases} \ln P_-(T_i^L), & \text{if } i = j \le n \\ \ln P_{link}(T_j^L|T_i^L) + 0.5[\ln P_+(T_i^L) + \ln P_+(T_j^L)] & \text{if } i,j \le n \text{ and } i \ne j \\ \ln P_{init}(T_j^L) + 0.5\ln P_+(T_j^L), & \text{if } i = j + n \\ \ln P_{term}(T_i^L) + 0.5\ln P_+(T_i^L), & \text{if } i + n = j \\ 0, & \text{if } i > n \text{ and } j > n \\ -\infty, & \text{otherwise} \end{cases} \tag{9}$$

in which $\ln P_+(T_i^L)$ is divided into two halves that are integrated into the two neighboring transition terms respectively.

As stated before, the MAP formulation in this paper takes false alarm hypotheses into account. In particular, this is represented by the diagonal components of the transition matrix: each one is set to be the logarithmic likelihood of the tracklet being a false alarm,

and the self-association of a tracklet is equivalent to rejecting it as a false alarm since it cannot be associated with any other tracklet, initialization or termination. Denoting $\Gamma^* = [\gamma_{ij}^*]_{2n\times 2n}$ as the optimal assignment matrix obtained by applying the Hungarian algorithm to the transition matrix $\mathbf{C}$, for each $\gamma_{ij}^* = 1$,

(1) if $i = j \leq n$, T_i^L is considered as a false alarm;
(2) if $i, j \leq n$ and $i \neq j$, link the tail of T_i^L to the head of T_j^L;
(3) if $i = j + n$, T_j^L is initialized as the head of the generated trajectory;
(4) if $i + n = j$, T_i^L is terminated as the tail of the generated trajectory.

In this way, we can compute $\mathcal{S}^*$ and its corresponding tracklet set $\mathcal{T}^{\mathcal{M}}$.

Implementation Details. The link probability between two tracklets is defined as the product of three components (appearance, motion and time):

$$P_{link}(T_j^L|T_i^L) = A_a(T_j^L|T_i^L)A_m(T_j^L|T_i^L)A_t(T_j^L|T_i^L). \tag{10}$$

To alleviate noises from inaccurate detections, for each input tracklet, a Kalman Filter is used to refine the positions and sizes of its detection responses and estimate their velocities. Color histograms of the detection responses are recomputed and integrated into a refined color histogram a_i^* for the tracklet by a RANSAC method.

The appearance affinity is defined by a Gaussian distribution:

$$A_a(T_j^L|T_i^L) = G(corr(a_i^*, a_j^*); 0, \sigma_c), \tag{11}$$

where $corr()$ calculates the correlation between a_i^* and a_j^*.

The motion affinity is defined as

$$A_m(T_j^L|T_i^L) = G\left(\mathbf{p}_i^{tail} + \mathbf{v}_i^{tail}\Delta t; \mathbf{p}_j^{head}, \mathbf{\Sigma}_{\Delta t}\right) G\left(\mathbf{p}_j^{head} - \mathbf{v}_j^{head}\Delta t; \mathbf{p}_i^{tail}, \mathbf{\Sigma}_{\Delta t}\right), \tag{12}$$

where Δt is the frame gap between the tail (i.e. the last detection response) of T_i^L and the head (i.e. the first detection response) of T_j^L; $\mathbf{p}_i^{head}$ (or $\mathbf{p}_i^{tail}$) and $\mathbf{v}_i^{head}$ (or $\mathbf{v}_i^{tail}$) are the refined position and estimated velocity of T_i^L at the head (or tail) (see Fig.2 for an illustration). The difference between the predicted position and the observed position is assumed to obey a Gaussian distribution.

The temporal affinity limits the maximum frame gap between two associated tracklets, and measures the probability of missed detections within the gap:

$$A_t(T_j^L|T_i^L) = \begin{cases} Z_\xi \alpha^{\Delta t - 1 - \omega}, & \text{if } \Delta t \in [1, \xi] \\ 0, & \text{otherwise} \end{cases}, \tag{13}$$

where α is the missed detection rate of the detector, ξ is an upper bound of frame gap, and Z_ξ is a normalization factor. Within the frame gap, ω is the number of frames in which the tracked object is occluded by other objects, and $\Delta t - 1 - \omega$ is the number of frames in which the tracked object is visible but missed by the detector. In practice, to compute ω, we interpolate detection responses within the frame gap (Fig.2) and check whether they are occluded by other objects by applying the occupancy map based occlusion reasoning method in [6] to $\mathcal{T}^{\mathcal{L}}$.

Initialization and termination probabilities of each tracklet are empirically set to be

$$P_{init}(T_i^L) = P_{term}(T_j^L) = Z_\xi \alpha^{\frac{1}{2}\xi}. \tag{14}$$

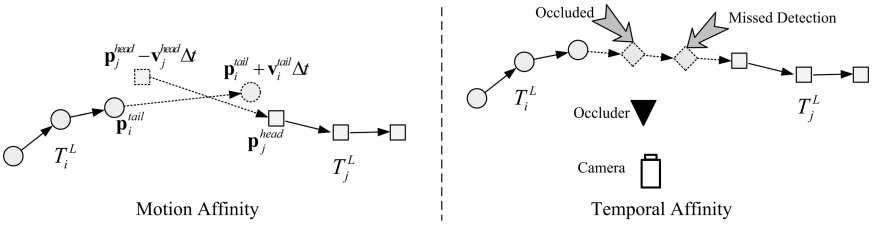

Fig. 2. Motion affinity and temporal affinity between two tracklets: the dashed circles and the dashed squares are predictions of T_i^L and T_j^L by their motion models, and the dashed diamonds are interpolated responses between the two tracklets

So far we have elaborated the first round at the middle level. In the following rounds, tracklets with longer frame gaps are associated by progressively increasing ξ.

2.3 High-Level Association

During the middle-level association, all tracklets have the same initialization/termination probabilities as there is no prior knowledge about entries and exits at that stage. This is equivalent to assuming uniform distributions of entries and exits in the scene. At the high level, an entry map and an exit map are inferred from $\mathcal{T}^{\mathcal{M}}$, which are used to specify the initialization/termination of each tracklet in the scene. In addition, a scene occluder map is also inferred from $\mathcal{T}^{\mathcal{M}}$ to revise the link probabilities. The three maps, as hidden variables, constitute a *scene structure model* in the high-level association. With a homography between the image plane and the ground plane, the scene structure model is estimated in the ground plane coordinates for better accuracy. We solve this coupled scene-estimation tracklet-association problem by an EM-like algorithm.

E-step. In the E-step, the probability distributions of entries, exits and scene occluders are calculated via Bayesian inference:

$$\frac{P\left(M_q(\overline{\mathbf{x}})=1|\overline{\mathbf{x}}\right)}{P\left(M_q(\overline{\mathbf{x}})=0|\overline{\mathbf{x}}\right)}=\frac{P\left(M_q(\overline{\mathbf{x}})=1\right)P\left(\overline{\mathbf{x}}|M_q(\overline{\mathbf{x}})=1\right)}{P\left(M_q(\overline{\mathbf{x}})=0\right)P\left(\overline{\mathbf{x}}|M_q(\overline{\mathbf{x}})=0\right)} \tag{15}$$

where $M_q(\overline{\mathbf{x}}) \in \{0,1\}$ is the indicator function for entries, exits or scene occluders ($q=\{en,ex,oc\}$) at position $\overline{\mathbf{x}}$ on the ground plane.

The positive and negative likelihoods are computed based on $\mathcal{T}^{\mathcal{M}}$. A tracklet T_k is used to propose a set of hypotheses for entries, exits and scene occluders as Fig.3.

The complete version of a tracklet T_k that includes missed detections, $\widetilde{T}_k=\{\mathbf{r}_{k_i}\}$, is obtained by filling the gaps between inconsecutive detection responses with interpolated ones. One observation is that an entry/exit is likely to be located close to the head/tail of a tracklet. Thus, two Gaussian hypotheses are proposed for the positive likelihoods of the entry and the exit respectively:

$$\begin{aligned} p_{en+}(\overline{\mathbf{x}}|T_k) &= G\left(\overline{\mathbf{x}};\ \overline{\mathbf{x}}_{en},\sigma_p\right),\ \ \overline{\mathbf{x}}_{en}=\overline{\mathbf{p}}_k^{head}-\overline{\mathbf{v}}_k^{head}\Delta t_m \\ p_{ex+}(\overline{\mathbf{x}}|T_k) &= G\left(\overline{\mathbf{x}};\ \overline{\mathbf{x}}_{ex},\sigma_p\right),\ \ \overline{\mathbf{x}}_{ex}=\overline{\mathbf{p}}_k^{tail}+\overline{\mathbf{v}}_k^{tail}\Delta t_m \end{aligned} \tag{16}$$

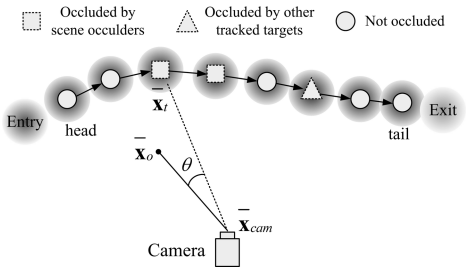

Fig. 3. Hypotheses for entries, exits and scene occluders proposed by a tracklet

where $\overline{\mathbf{p}}_k^{head}/\overline{\mathbf{p}}_k^{tail}$ and $\overline{\mathbf{v}}_k^{head}/\overline{\mathbf{v}}_k^{tail}$ are the estimated position and velocity at the head/tail of $\widetilde{T}_k^H$ by the Kalman Filter, and Δt_m is a short time span for predicting the positions of the entry and the exit. Another observation is that entries and exits are unlikely to be close to the passed region of a tracklet. Therefore, a mixture of Gaussian hypotheses is adopted to model their negative likelihoods:

$$p_{en-}(\overline{\mathbf{x}}|T_k) = p_{ex-}(\overline{\mathbf{x}}|T_k) = \frac{1}{|\widetilde{T}_k|} \sum_{\mathbf{r}_{k_i} \in \widetilde{T}_k} G\left(\overline{\mathbf{x}};\ \overline{\mathbf{x}}_{k_i}, \sigma_p\right) \tag{17}$$

where $\overline{\mathbf{x}}_{k_i}$ is the position of response $\mathbf{r}_{k_i}$.

Estimation of the scene occluder map focuses on the area between the camera and the tracklet. Given the camera position $\overline{\mathbf{x}}_{cam}$ on the ground plane, an angle-based Gaussian distribution is defined to model the occlusion relationship between a target at position $\overline{\mathbf{x}}_t$ and a scene occluder at position $\overline{\mathbf{x}}_o$:

$$\widetilde{G}_{angle}(\overline{\mathbf{x}}_o, \overline{\mathbf{x}}_t, \overline{\mathbf{x}}_{cam}) = Z_a G\left(angle(\overline{\mathbf{x}}_o - \overline{\mathbf{x}}_{cam}, \overline{\mathbf{x}}_t - \overline{\mathbf{x}}_{cam}); 0, \sigma_a\right) \tag{18}$$

where $angle()$ computes the view angle difference between the target and the occluder (i.e. θ shown in Fig.3), and Z_a is a normalization factor. The probability is maximized when the occluder lies on the line segment between the target and the camera.

An interpolated response of a tracklet is not detected by the detector. If it is not occluded by any other tracked target, it might be occluded by a scene occluder. Conversely, with high probability, there are no scene occluders lying between the camera and the detection responses. Based on these observations, a tracklet gives positive and negative likelihoods of scene occluders by

$$\begin{aligned} p_{oc+}(\overline{\mathbf{x}}|T_k) &= \frac{1}{|O_k|} \sum_{\mathbf{r}_{k_i} \in O_k} \widetilde{G}_{angle}\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}_{k_i}, \overline{\mathbf{x}}_{cam}\right), \\ p_{oc-}(\overline{\mathbf{x}}|T_k) &= \frac{1}{|T_k|} \sum_{\mathbf{r}_{k_i} \in T_k} \widetilde{G}_{angle}\left(\overline{\mathbf{x}}, \overline{\mathbf{x}}_{k_i}, \overline{\mathbf{x}}_{cam}\right) \end{aligned} \tag{19}$$

where T_k is the original tracklet whose responses are all detected, and O_k is a subset of the complete tracklet $\widetilde{T}_k$, consisting of the interpolated responses that are not occluded

by any other tracked target. Again the occlusion reasoning method in [6] is adopted here to decide the occlusion type of an interpolated response. Notice that interpolated responses that are occluded by other tracked targets make no contributions to the estimation of scene occluders.

Considering the hypotheses proposed by each tracklet as a set of i.i.d. samples extracted from the true likelihood distributions, we approximate the likelihoods in Equ.15 by the mixture of all hypotheses.

$$P\left(\overline{\mathbf{x}}|M_q(\overline{\mathbf{x}})=1\right)=\frac{1}{|\mathcal{T}^{\mathcal{M}}|}\sum_{T_k\in\mathcal{T}^{\mathcal{M}}}p_{q+}(\overline{\mathbf{x}}|T_k),$$
$$P\left(\overline{\mathbf{x}}|M_q(\overline{\mathbf{x}})=0\right)=\frac{1}{|\mathcal{T}^{\mathcal{M}}|}\sum_{T_k\in\mathcal{T}^{\mathcal{M}}}p_{q-}(\overline{\mathbf{x}}|T_k),\quad q=\{en,ex,oc\}. \tag{20}$$

Eventually, the posterior probabilities in Equ.15 can be computed with their corresponding predefined prior probabilities.

M-step. In the M-step, the tracklets in $\mathcal{T}^{\mathcal{M}}$ are further associated to form even longer ones. Similar to the middle level, the association problem is formulated as a MAP problem and solved by the Hungarian algorithm. However, based on the scene structure model obtained from the E-step, the initialization and termination probabilities (Equ.14) of each tracklet are recomputed as

$$P_{init}(T_k)=Z_\xi\alpha^{\min(\frac{1}{2}\xi,\Delta t_{init})},\quad P_{term}(T_k)=Z_\xi\alpha^{\min(\frac{1}{2}\xi,\Delta t_{term})}. \tag{21}$$

Δt_{init}(or Δt_{term}) is the frame number of missed detection part between the head (or tail) of T_k to the nearest entry (or exit):

$$\Delta t_{init}=\inf\left\{\Delta t:\ P\left(M_{en}(\overline{\mathbf{x}})=1|\overline{\mathbf{x}}=\overline{\mathbf{x}}_k^{head}-\overline{\mathbf{v}}_k^{head}\Delta t\right)>0.5\right\},$$
$$\Delta t_{term}=\inf\left\{\Delta t:\ P\left(M_{ex}(\overline{\mathbf{x}})=1|\overline{\mathbf{x}}=\overline{\mathbf{x}}_k^{tail}+\overline{\mathbf{v}}_k^{tail}\Delta t\right)>0.5\right\}. \tag{22}$$

where $\inf\{\}$ is the infimum of a set. Moreover, an interpolated response $\mathbf{r}_0$ at position $\overline{\mathbf{x}}_0$ is considered to be occluded by scene occluders if and only if

$$\max_{\alpha\in[0,1]}\left[P\left(M_{oc}(\overline{\mathbf{x}})=1|\overline{\mathbf{x}}=\alpha\overline{\mathbf{x}}_0+(1-\alpha)\overline{\mathbf{x}}_{cam}\right)\right]>0.5. \tag{23}$$

This is used to reivse the temporal affinity in Equ.13 by considering occlusions by scene occluders when counting the occluded frame number ω.

The scene structure model helps explain three important events for tracklet association: entering the scene, exiting the scene and being occluded by scene occluders. This greatly reduces the ambiguity of associating tracklets with long frame gaps. Compared to the parametric sources/sink model in [7] and the semi-automatic scene understanding module in [17], our approach is nonparametric and fully automatic. It directly calculates the posterior probability maps by Bayesian inference based on object tracklets.

Table.2 summarizes the overall algorithm of our hierarchical framework.

Table 2. The overall algorithm of hierarchical association framework

0) (Given): the detection response set $\mathcal{R}$, upper bounds of frame gap $\{\xi_i | \ \xi_i < \xi_{i+1}\}$ and the number of iterations in the middle-level association D

1) (Low-level association): obtain tracklet set $\mathcal{T}^{\mathcal{L}}$ by the direct link method.

2) (Middle-level association):

- Initialize: $\mathcal{T}^{\mathcal{M}} = \mathcal{T}^{\mathcal{L}}$.
- For $i = 1$ to D
  - ◇ (Affinity Revision): for each tracklet in $\mathcal{T}^{\mathcal{M}}$, obtain a motion model by Kalman filter and a refined appearance model by a RANSAC method; for each frame, compute an occupancy map according to $\mathcal{T}^{\mathcal{M}}$.
  - ◇ (Association): calculate the transition matrix $\mathbf{C}$ with ξ_i, obtain the optimal tracklet association set $\mathcal{S}^*$ and the corresponding $\mathcal{T}^*$. Set $\mathcal{T}^{\mathcal{M}} = \mathcal{T}^*$.

3) (High-level association):

- (E-step): estimate the scene structure model from $\mathcal{T}^{\mathcal{M}}$.
- (M-step): based on the scene structure model, repeat the iterative process in the middel-level association once with $\xi = \xi_{D+1}$ to obtain trajectory set $\mathcal{T}^{\mathcal{H}}$.

4) (Output): the complete trajectory set $\widetilde{\mathcal{T}}^{\mathcal{H}}$ by filling the frame gaps in $\mathcal{T}^{\mathcal{H}}$.

3 Experimental Results

We apply our hierarchical association framework to the multiple pedestrian tracking problem. In all experiments, the number of iterations D is set to be 3, and the upper bounds of frame gap ξ_i are set to be $8, 32, 128$ for the middle level and 256 for the high level. We evaluate our method on two public video corpora: the CAVIAR set [19] and the i-LIDS AVSS AB set [20]. The CAVIAR set contains 26 videos captured in a corridor, and its ground truth contains 235 trajectories. The i-LIDS AVSS AB set contains three videos captured in a subway station, and its ground truth includes 135 trajectories. Both data sets are challenging due to heavy occlusions. We learn our pedestrian detector by the method in [6], and none of the videos in these two test sets are used for training.

Evaluation Metric. We adopt the metrics in the CLEAR evaluation [18] and use an automatic scoring software provided by the organizer. The three adopted metrics are:

(1) MOTA: Multiple Object Tracking Accuracy, calculated from the number of false alarms, missed detections, and identity switches;

(2) FGTIM: Fraction of Ground Truth Instances Missed; and

(3) FAPF: False Alarm Per Frame.

For the first score, higher is better; for the last two scores, lower is better.

For comparison, we also evaluate the method in [6] on the two sets. Table.3 lists the scores. Among them, "Wu & Nevatia's [6]" and "Our high level" are the final results of the two methods; "Our low level" and "Our middle level round i" ($i = \{1, 2, 3\}$) are the intermediate results obtained in our method, which are used to demonstrate the progressive improvement achieved by the hierarchical association framework. In addition, to show the advantage of the iterative process at the middle level, a simplified version of middle-level association with only one round, denoted as "Our middle level #", is evaluated. To show the benefit from the E-step at the high level, we evaluate a degenerate

Table 3. Evaluation results on CAVIAR and i-LIDS with CLEAR metric

	CAVIAR			i-LIDS		
	MOTA	FGTIM	FAPF	MOTA	FGTIM	FAPF
Wu & Nevatia's [6]	0.537	0.470	0.012	0.553	0.370	0.228
Our high level	0.800	0.200	0.025	0.684	0.290	0.137
Our low level	0.540	0.338	0.395	0.475	0.507	0.080
Our middle level round 1	0.627	0.332	0.141	0.490	0.507	0.042
Our middle level round 2	0.694	0.292	0.064	0.547	0.459	0.024
Our middle level round 3	0.759	0.235	0.032	0.640	0.358	0.059
Our middle level #	0.705	0.263	0.118	0.592	0.401	0.060
Our high level #	0.771	0.223	0.041	0.656	0.343	0.062

version of the high-level association (denoted as "Our high level #"), which replaces the estimated scene structure model with a general one. The homography between the image plane and the ground plane is obtained by manually labeling several points on the ground. This could also be easily derived from camera parameters.

CAVIAR Set. As shown in Table.3, from the low level to the high level, the MOTA score of our method is progressively increased, while the FGTIM score and the FAPF score are gradually decreased. The low-level association has the lowest detection rate and the highest false alarm rate since it just links detection responses in consecutive frames and doesn't take false alarm hypotheses into account. As the hierarchical association proceeds, the upper bound of frame gap ξ increases from 8 to 256. On one hand, it enables the algorithm to associate strong tracklets with long frame gaps, so that more and more missed detections can be recovered; on the other hand, it decreases the initialization and termination probabilities of each tracklet (see Equ.14), so that weak tracklets are likely to be rejected as false alarms. Compared to Wu & Nevatia's method [6], ours achieves a comparably low FAPF score, a much lower FGTIM score, and a much higher MOTA score. The first row (input detection responses) and the second row (tracking results) of Fig.4 demonstrate the improvements achieved by our method: 1) two false alarms in frame 133 are rejected by the proposed false alarm hypotheses in the MAP formulation; 2) three missed detections in frame 233 are recovered by associating long-gap tracklets; 3) an inaccurate response in frame 407 is corrected. The third row is the tracking result of another video in the CAVIAR set, and more example results are submitted as supplementary materials. The experimental results show that our method can overcome some deficiencies of the pedestrian detector as shown in Fig.1.

Compared to the iterative association approach ("Our middle level round 3") at the middle level, the simplified one ("Our middle level #") is inferior in terms of all three scores. Without the estimated specific scene structure model, the degenerate high-level association ("Our high level #") can not perform as well as the original one ("Our high level"). The comparisons justify the use of the iterative association approach and the specific scene structure model in our method.

i-LIDS Set. Similar to the results on the CAVIAR set, our method achieves big improvement for MOTA and FGTIM on the i-LIDS set. However, the FAPF score is not

Fig. 4. Tracking results of our method on the CAVIAR set: the first row and the second row are the input detection responses and the corresponding tracking results for a video; the third row is the tracking result for another video

monotonously reduced: it drops to the minimum at the second round of the middle level and rises afterward. We attribute this to the extremely heavy inter-occlusions in this data set (as shown in the second row of Fig.5): when a train stops at the station, many pedestrians rush in and out of the carriage in a short time. Tracklets of pedestrians in such crowded environments are highly fragmented and noisy. Associating these tracklets with long frame gaps is risky, since it is difficult to distinguish too many disordered targets with similar appearance based on color and motion. However, as the tradeoff between the detection rate and the false alarm rate, the MOTA score keeps increasing, and the final results of our method outperform those of Wu & Nevatia's [6] for all three scores. Again, the difference between the two modified methods ("Our middle level #" and "Our high level #") and their original versions ("Our middle level round 3" and "Our high level") justify the use of the iterative association approach and the specific scene structure model. Fig.5 shows some tracking results of our method on this set.

Scene Structure Models and Tracking Speed. In practice, the three probability maps in the scene structure model are implemented as discrete grids for efficiency. To visualize the estimated scene structure model, we binarize every probability map with a threshold of 0.5 as shown in Fig.6. It can be seen that our Bayesian inference approach in the E-step of the high level is capable of effectively reconstructing the crucial environment elements for the final association in the M-step. In particular, our method successfully locates the pillars as occluders in both scenes, which is important for the recovery of trajectories occluded by the pillar for a long time. The first row of Fig.5

Fig. 5. Tracking results of our method on i-LIDS set

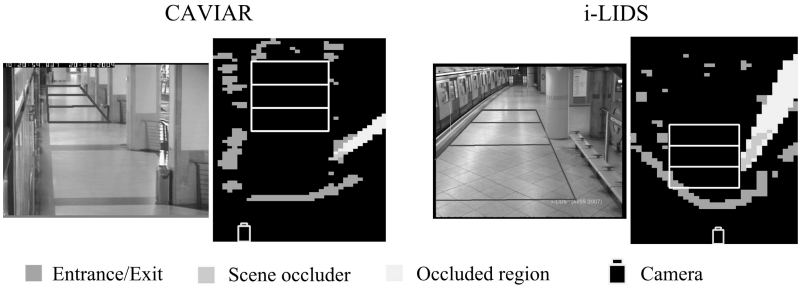

Fig. 6. Estimated scene structure models for CAVIAR and i-LIDS: the entry map and the exit map are both painted in red as they are mostly overlapped; the occluded region (in yellow) is inferred from the positions of camera (at the bottom) and scene occluders (in green); the three rectangles of each image indicate the ground plane coordinates.

shows an example from the i-LIDS set, in which the trajectory 15 is not broken although it is occluded by the pillar for more than 200 frames (i.e. 8 seconds).

For computational efficiency, a sliding window technique is used at the middle level to reduce the size of transition matrix for the Hungarian algorithm. Given the detection results, the speed of our method is about 50 FPS on a 3.0G Hz PC.

4 Conclusion

In this paper, we present a robust hierarchical association framework for the multiple object tracking problem. Experimental results on two challenging data sets show that the proposed method significantly improves the tracking performance by effectively associating tracklets with inaccurate detection responses and long-time occlusions. This hierarchical framework is a general approach, and other affinity measures or optimization methods can be easily integrated into this framework.

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

References

1. Comaniciu, D., Ramesh, V., Meer, P.: The Variable Bandwidth Mean Shift and Data-Driven Scale Selection. In: ICCV (2001)
2. Tomasi, C., Kanade, T.: Detection and Tracking of Point Features. Technical Report CMU-CS-91-132, Carnegie Mellon University (1991)
3. Zhao, T., Nevatia, R.: Tracking Multiple Humans in Complex Situations. IEEE trans. on PAMI 26(9), 1208–1221 (2004)
4. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: CVPR (2001)
5. Zhu, Q., Avidan, S., Yeh, M.C., Cheng, K.T.: Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: CVPR (2006)
6. Wu, B., Nevatia, R.: Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. International Journal of Computer Vision (2007)
7. Stauffer, C.: Estimating tracking sources and sinks. In: IEEE Workshop on Event Mining in Video (2003)
8. Perera, A.G.A., Srinivas, C., Hoogs, A., Brooksby, G., Hu, W.: Multi-Object Tracking Through Simultaneous Long Occlusions and Split-Merge Conditions. In: CVPR (2006)
9. Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Tracking in Low Frame Rate Video: A Cascade Particle Filter with Discriminative Observers of Different Lifespans. In: CVPR (2007)
10. Kuhn, H.W.: The Hungarian method for the assignment problem. Naval Research Logistics Quarterly 2, 83–87 (1955)
11. Reid, D.: An algorithm for tracking multiple targets. IEEE Trans. Automatic Control 24, 843–854 (1979)
12. Fortmann, T., Shalom, Y.B., Scheffe, M.: Sonar tracking of multiple targets using joint probabilistic data association. IEEE J. Oceanic Engineering 8, 173–184 (1983)
13. Leibe, B., Schindler, K., Gool, L.V.: Coupled Detection and Trajectory Estimation for Multi-Object Tracking. In: ICCV (2007)
14. Okuma, K., Taleghani, A., Freitas, N.D., Little, J.J., Lowe, D.G.: A Boosted Particle Filter: Multitarget Detection and Tracking. In: ECCV (2004)
15. Yu, Q., Medioni, G., Cohen, I.: Multiple target tracking using spatio-temporal markov chain monte carlo data association. In: CVPR (2007)
16. Singh, V.K., Wu, B., Nevatia, R.: Pedestrian Tracking by Associating Tracklets using Detection Residuals. In: WMVC (2008)
17. Kaucic, R., Perera, A.G.A., Brooksby, G., Kaufhold, J., Hoogs, A.: A Unified Framework for Tracking through Occlusions and across Sensor Gaps. In: CVPR (2005)
18. CLEAREVAL, `http://www.clear-evaluation.org/`
19. CAVIAR, `http://homepages.inf.ed.ac.uk/rbf/CAVIAR/`
20. iLIDS, `http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html`