# Robust Multi-Human Tracking by Detection Update using Reliable Temporal Information

Lu Wang, Qingxu Deng and Mingxing Jia
*College of Information Science and Engineering, Northeastern University, Shenyang, China*

Keywords: Multi-Target Tracking, Data Association, Detection Update, Video Surveillance.

Abstract: In this paper, we present a multiple human tracking approach that takes the single frame human detection results as input, and associates them hierarchically to form trajectories while improving the original detection results by making use of reliable temporal information. It works by first forming tracklets, from which reliable temporal information can be extracted, and then refining the detection responses inside the tracklets. After that, local conservative tracklets association is performed and reliable temporal information is propagated across tracklets. The global tracklet association is done lastly to resolve association ambiguities. Comparison with two state-of-the-art approaches demonstrates the effectiveness of the proposed approach.



## 1 INTRODUCTION

Object tracking in video surveillance aims at extracting objects' spatial-temporal information, which is mandatory for higher level activity recognition. However, it is not trivial due to difficulties such as low figure-ground contrast, changes in object appearance over time, abrupt motions, and etc. Multiple object tracking is even more challenging as inter-object occlusions exist prevalently.

Tracking by detection is effective for solving the multiple object tracking problem and therefore has been widely applied (Wu, 2007; Huang, 2008, Breitenstein, 2011). It usually consists of two steps: The first one is time-independent object detection and the second is detection responses association temporally based on appearance similarity, motion consistency, etc. Compared to traditional visual tracking (Rasmussen, 2001; Comaniciu 2003; Wang, 2010), tracking by detection can effectively avoid the drifting problem caused by accumulated tracking error. In addition, it is robust to occasional detection failures, i.e. isolated false alarms or missed detections are less likely to lead to tracking failures.

As a good compromise between association accuracy and computation complexity, tracklet (i.e. track fragment) based approaches (Stauffer, 2003; Perera, 2006; Huang, 2008) have become more and more popular. In these approaches, tracklets are first generated by conservative linking of detections of consecutive frames, which helps reduce the possible linking space significantly. Then, given the affinities between potentially linkable tracklets, the association problem is typically solved by the Hungarian algorithm (Munkres, 1957). For example, Stauffer associates tracklets using the Hungarian algorithm with an extended transition matrix that considers the likelihood that each tracklet being the initialization and termination of trajectories (Stauffer, 2003). This approach performs iterative tracklet association and scene entrances/exits estimation using expectation maximization . Perera *et al*. adapt the Hungarian algorithm to deal with the merging and splitting of tracklets in multiple object tracking when long time occlusion exists (Perera, 2006). Both Stauffer and Perera's approaches define the tracklet affinity only once, which may not be accurate enough due to the errors introduced by inaccurate localization in the detection phase. Huang *et al*. propose a hierarchical data association strategy, in which tracklet affinities are refined whenever new tracklets are formed during the progressive tracklet linking procedure (Huang, 2008). To further increase the robustness of the affinity measures, a few approaches have been recently proposed. For example, Li *et al*. propose a HybridBoost algorithm to learn the affinity models between two tracklets (Li, 2009). Kuo *et al*. propose global on-line

Wang L., Deng Q. and Jia M..
Robust Multi-Human Tracking by Detection Update using Reliable Temporal Information.
DOI: 10.5220/0004733203870396
In *Proceedings of the 9th International Conference on Computer Vision Theory and Applications* (VISAPP-2014), pages 387-396
ISBN: 978-989-758-009-3

discriminative appearance models, where descriptors are pre-defined (Kuo, 2010), and later they propose to use automatic important feature selection by learning from a large number of local image descriptors (Kuo, 2011). Yang *et al*. propose to learn non-linear motion patterns to better explain direction changes (Yang, 2012).

In this paper, we present a tracklet based data association approach for multiple human tracking in surveillance scenarios, with the assumptions that the camera is static, people walk on a ground plane and camera parameters can be obtained. Unlike most of the previous data association works that only consider how to ensure correct linking, we also attempt to improve the detections when reliable temporal information can be obtained. To this end, we first generate tracklets by conservative linking of detections, and extract the appearance, size and position information of those reliable detections that show high temporal and spatial consistency. Then the extracted information is propagated to detections within the tracklets by refining the detections' shape models. After that, local conservative tracklet association based on the Hungarian algorithm is performed so that reliable temporal information can be further propagated. The iteration stops when there are no new detection updates or new tracklet association. Finally, the Hungarian algorithm is applied globally to resolve ambiguous situations and . The whole process ends when neither new updates nor association can be performed. The outputs of the approach are the updated detection responses as well as the associated trajectories.

Our proposed approach is most related to Huang *et al.*'s approach (Huang, 2008), where data association is formulated as a Maximum a Posteriori (MAP) problem and solved by the Hungarian algorithm. We made necessary improvements over that approach for more robust performance. The first one is that reliable temporal information is extracted to improve the quality of detections and tracklets. The second one is that we propose a local tracklet association procedure before global association, which is more conservative and less likely to make errors. The third one is that we use the reliable temporal information to recover missed head or tail parts of tracklets, bringing in the advantages of shortening gaps between tracklets, hence enabling associations to be made more robustly, and making the resulting tracks more complete. The fourth one is that we detect tracklets that may violate the 1st-order Markov Chain assumption and approximate the 2nd-order Markov Chain on them.

In summary, our main contributions are three-fold:

1) Improving the accuracy of human detection by using reliable temporal information;
2) A new iterative hierarchical data association framework;
3) When perform global data association, explicitly detecting tracklets that may violate the 1st-order Markov Chain assumption and approximate the 2nd-order Markov Chain on them.

## 2 THE PROPOSED APPROACH

In this section, we will introduce the detail of the proposed association approach. The diagram of the approach is illustrated in Figure 1.

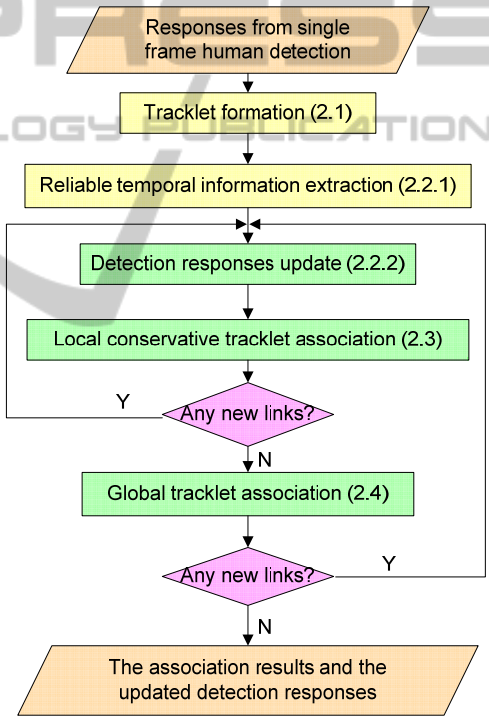



Figure 1: The diagram of the proposed data association approach.

### 2.1 Human Detection and Tracklet Formation

The single frame human detection result is firstly obtained using a crowd detection approach (e.g. Wu, 2007). Then model fitting, as proposed in Wang, 2012, is applied to find the best matched 3D shape model for each detection response **r,** with the model parameters being the 3D location **z**, the orientation

$o$, the size $s$ and the leg posture $pose$.

Given all the detection responses $\mathcal{R}=\{\mathbf{r}_i\}$ and the corresponding model parameters, the detections are firstly associated conservatively to form the tracklets, and the affinity $A(\mathbf{r}_i, \mathbf{r}_j)$ between any two detection responses $\mathbf{r}_i$ and $\mathbf{r}_j$ is defined as:

$$A(\mathbf{r}_i, \mathbf{r}_j)=\begin{cases} A_{app}(\mathbf{r}_i, \mathbf{r}_j)A_{pos}(\mathbf{r}_i, \mathbf{r}_j) & \text{if } f_j - f_i = 1 \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $A_{app}(\mathbf{r}_i, \mathbf{r}_j)$ and $A_{pos}(\mathbf{r}_i, \mathbf{r}_j)$ represents the appearance similarity and position proximity respectively, and $f_i$ denotes $\mathbf{r}_i$'s frame index.

As the human model is part based, to make the appearance model more discriminative, we define a three-part appearance model $\mathbf{a} = \{\mathbf{a}_{pt} | pt=h, t, l\}$ for each detection, where $pt$ denotes the body part, $h$, $t$ and $l$ represent head, torso and legs respectively, and each $\mathbf{a}_{pt}$ is an $8\times8\times8$ RGB color histogram. The appearance affinity is calculated as

$$A_{app}(\mathbf{r}_i, \mathbf{r}_j)=\frac{\sum_{pt=\{h,t,l\}} w_{pt}\min(v_{i,pt}, v_{j,pt})BC(\mathbf{a}_{i,pt}, \mathbf{a}_{j,pt})}{\sum_{pt=\{h,t,l\}} w_{pt}\min(v_{i,pt}, v_{j,pt})}, \quad (2)$$

where $v_{i,pt}$ is the visible ratio of part $pt$ of $\mathbf{r}_i$, $BC$ calculates the Bhattacharyya coefficient of two histograms, and $w_{pt}$ is the weight for part $pt$. For a human object, as the head and torso are more accurately described by the model than legs, they should have higher weights than legs. Therefore, we set $w_h=w_t=0.4$ and $w_l=0.2$. Slight changes of these weights would make no obvious difference.

The position proximity of two detection responses is defined in terms of the distance $d$ traversed by a human at a high speed within one time step:

$$A_{pos}(\mathbf{r}_i, \mathbf{r}_j)=\begin{cases} 1 & \text{if } |\mathbf{z}_i - \mathbf{z}_j| \le d, \\ G(\max(|\mathbf{z}_i - \mathbf{z}_j| - d, 0); 0, \sigma_z) & \text{otherwise} \end{cases} \quad (3)$$

where $G(\cdot; x_0, \sigma_x)$ represents the Gaussian distribution with mean $x_0$ and standard deviation $\sigma_x$. In our experiment, we set $d$ to be 0.7 meters when the time step is 0.4 seconds.

Having the affinity values, the two-threshold strategy as presented in Huang, 2008 is used to generate the tracklets, i.e. 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.

## 2.2 Detection Responses Update

Given the tracklets, reliable temporal information is extracted from them and used to refine related detections by means of model matching. By reliable temporal information, we mean that the appearance, position and size information of the corresponding detection is accurate enough so that these pieces of information can be used to guide the update of detections of same identity in neighboring frames.

### 2.2.1 Reliable Temporal Information Extraction

To extract reliable temporal information, we first look for the reliable detections. Specifically, we detect reliable detections according to the following criteria:

a) The detection's head contour is well aligned with the foreground contour; and
b) The detection has high appearance, head position and feet position affinities with its adjacent detections; and
c) There are at least three consecutive detections from the same tracklet that satisfy a) and b) simultaneously (We choose three to avoid the coincidental satisfaction of condition a) and b).).

Figure 2 illustrates some reliable detection responses found by the above criteria.

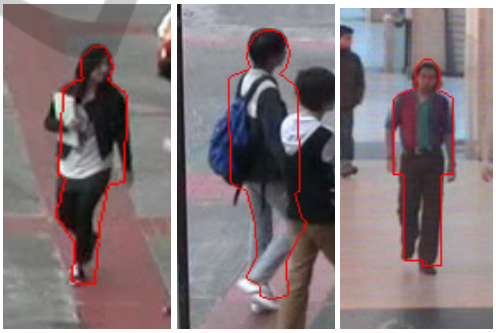


Figure 2: Illustration of reliable detections.

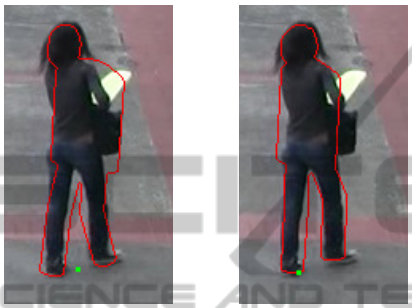
Next, we refine the models of unoccluded reliable detections by using the temporal information provided by the tracklets. Those occluded ones will be refined when their occluding objects have been refined. The best matched model $M$ for a reliable detection is selected as

$$M = \max_{m(pose,s,o,\mathbf{p}_0)} \frac{G(s; s_0, \sigma_s)G(o; o_0, \sigma_o)}{L_s(B(m(pose, s, o, \mathbf{p}_0)))}, \quad (4)$$

where $m$ represents the 3D model; $B(m)$ is the model's boundary on the image; $s_0$ is the average size of the reliable detections within the corresponding tracklet; $o_0$ is the detection's tangential direction in the tracklet; $L_s$ is the shape

likelihood measuring how well the model matches with the image edges (see Wang, 2012 for details); $\mathbf{p}_0$ is the original model's head position on the image, which is assumed to be accurate and needs not to be refined. Figure 3 shows an example of the best matched model before and after using the temporal information, where the original wrong orientation estimation has been corrected.

Having the updated models for the reliable detections, we then extract the corresponding appearance, size and position information from them and propagate it to the other frames.



(a) (b)

Figure 3: Illustration of the best matched models without and with the temporal information. (a) Without temporal information; (b) With temporal information.

### 2.2.2 Temporal Information Propagation

For an unoccluded detection adjacent to an updated detection in the same tracklet, we refine it by using the reliable temporal information. An **unoccluded** detection means that the person corresponding to the detection is fully visible, or occluded by the image border, or a human object whose occluding human objects have already been updated. Unlike reliable detections, the original head position estimation for an ordinary detection may not be correct. Therefore, we first predict its head and feet positions separately using the temporal information. If the predicted position is quite near to the original position, the original position is taken to be correct; otherwise, both the original and the predicted positions are checked. Specifically:

(a) If the original head position $\mathbf{p}_0$ is considered to be correct, we do model matching by

$$M = \max_{m(pose,s,o,\mathbf{p}_0)} G(s;s_0,\sigma_s)G(o;o_0,\sigma_o) \\ A_{app}(m(pose,s,o,\mathbf{p}_0),\mathbf{r}_{ref})L_s(B(m(pose,s,o,\mathbf{p}_0))) \quad (5)$$

where $\mathbf{r}_{ref}$ is the referenced detection response in the adjacent frame. The difference between (4) and (5) is the appearance affinity term $A_{app}$, as in the current stage we have an appearance model to refer to.

(b) Otherwise, if the feet position is taken to be correct, the head position $\mathbf{p}_0$ is first searched using the following equation

$$\mathbf{p}_0 = \max_{\mathbf{p}} G(s;s_0,\sigma_s)G(o;o_0,\sigma_0) \\ A_{app}(ubm(s,o,\mathbf{p}),\mathbf{r}_{ref})L_s(B(ubm(s,o,\mathbf{p}))) \quad (6)$$

where *ubm* represents the upper body model, which is used to avoid the high computational cost of searching for the optimal leg pose. Then model matching is performed according to (5).

(c) If neither the head position nor the feet position is correct, we search the head position using the original and the predicted feet positions separately (both the original and predicted positions are considered here to take into account of sudden motion changes) as in step (b). Then the searched head position with higher upper body appearance affinity to the reference appearance model is taken to be correct and model matching is performed on it as in step (a).

After obtaining the best matched model, the detection's appearance is updated. To deal with occlusion, the appearance model **a** is renewed by

$$\mathbf{a} = \alpha\mathbf{v}\mathbf{a}_{model} + (1-\alpha\mathbf{v})\mathbf{a}_{ref}, \quad (7)$$

where $\mathbf{a}_{model}$ is the appearance model of the newly obtained shape model; $\mathbf{a}_{ref}$ is the referenced appearance model; $\mathbf{v}=\{v_{pt}|\ pt=h,\ t,\ l\}$ and $v_{pt}$ is the visible ratio of part *pt*; $\alpha$ is the smoothing factor, which helps avoid large changes of the appearance model caused by incorrect detection update, and is set to be 0.2. A body part *pt* is not updated if $v_{pt}$ is less than 0.5, as in such situation the part is considered as severely occluded and hence unreliable.

### 2.2.3 Occlusion Order Determination

As our approach requires that a detection be updated only when it becomes unoccluded. However, the occlusion order obtained from the original detection result may contain inaccuracy and hence needs to be properly dealt with.

For two detections whose heads are at the similar horizontal level in the image and whose torsos intersect for only a small percentage (5% is used in our experiment), we consider they are not mutually occluded. In addition, we assume that, if it can be definitely determined that a detection *A* occludes another detection *B* in one frame *f*, it is impossible that *B* can occlude *A* in frame *f*-1 and frame *f*+1, because normally two persons could not change the

occlusion order within a very short time interval.

Furthermore, false positives (FPs), may introduce problem when propagating the reliable temporal information, because FPs may occlude some true detections and FPs are very unlikely to be updated as it cannot link to any reliable detections. Therefore, we collect the detections that cannot be linked to any other detection responses as candidate FPs. By doing so, a detection response is allowed to be updated if it is only occluded by a candidate FP. In case that a candidate FP is found to have a high affinity to an updated detection, it is not taken as an FP anymore.

## 2.3 Conservative Local Data Association

After all the possible detection updates have been made inside each tracklet, the tracklets can be associated. However, as only a small portion of detections may have been updated at this stage, the tracklets' link probability may still contain many inaccuracies, making global association of tracklets too risky to perform at this time. Therefore, we introduce an intermediate tracklets association step, namely local conservative Hungarian linking, which only aims at performing association for tracklets that exhibit high link probability and low ambiguity, and at the same time have no gaps in between. In addition, if an end of a tracklet has been updated but has no other tracklets to link to, we infer that some object might be missed at the detection stage and use the detection update method as a detector to recover the missed detections. For the other more ambiguous connections, we leave them to the later global Hungarian linking.

From here on, we use double subscripts to represent the quantities corresponding to the detections of a tracklet, with the first denoting the index of the tracklet and the second denoting the detection's index inside the tracklet. For example, the detections of a tracklet $T_i$ is denoted as $\mathbf{r}_{i,k}$, where $k=\{1, 2, \ldots,|T_i|\}$. We also denote by $\mathcal{T}_f^{end}=\{T_i\}$ the set of all the tracklets that end at frame $f$ and $\mathcal{T}_{f+1}^{start}=\{T_j\}$ the set of all the tracklets that start at frame $f$+1.

### 2.3.1 Local Affinity Definition

In conservative tracklets association, the link probability between two tracklets $T_i \in \mathcal{T}_f^{end}$ and $T_j \in \mathcal{T}_{f+1}^{start}$ is defined as

$$P_{link\_local}(T_i, T_j)= P_{app}(T_i, T_j)P_{local\_motion}(T_i, T_j). \quad (8)$$

$P_{app}(T_i, T_j)$ calculates the affinity according to (2) between the average appearance model of the last three detections of $T_i$ and that of the first three detections of $T_j$

Denoting $T_i$'s predicted model at its rear end as $\mathbf{r}_{i,|T_i|+1}$ and $T_j$'s predicted model at its front end as $\mathbf{r}_{j,0}$, $P_{local\_motion}(T_i, T_j)$ is calculated by the average intersection ratios of the predicted model and the corresponding detection.

### 2.3.2 Local Data Association

For each frame $f$, we apply the Hungarian algorithm on the resulting link probability matrix to obtain the tracklets correspondence. As the correspondence is only calculated locally, we cannot accept all the correspondences but only those reliable ones. Therefore, we firstly accept the links with high reliability using the two-threshold strategy; we then accept the correspondences with relatively high link probability and at the same time the tracklets contained are not linkable to any other tracklet that is not involved in any correspondences. The latter condition ensures that the accepted correspondences introduce no controversial association.

### 2.3.3 Missed Detection Recovery

In addition to the accepted correspondences, there is another situation we can deal with, i.e. if a tracklet has one end detection updated but is not linkable to any other tracklet, while that end is not at the image border or in the scene occluder areas, we are sure that the corresponding object is missing. In this case, we use the procedure stated in Section 2.2.2 to detect the missed object, with the difference being that we only have the predicted position. To avoid drifting, we accept the detection only if it has a high appearance affinity to the reference model, the shape matching score is high and it does not overlap significantly with other existed detections in the frame. Figure 4 shows an example of recovered missed detections.

After the local association, reliable temporal information can be propagated again, and new association can be made when more detections have been updated. The iteration continues until there are no new detection updates or local data associations.

## 2.4 Global Data Association usingthe Hungarian Algorithm

When no further update or association can be made, we resort to the global Hungarian tracklets

(a) (b)

Figure 4: Illustration of the result of tracklet extension. (a) The single frame detection result; (b) the detections recovered by our approach.

association, where gaps are allowed between associated tracklets.

### 2.4.1 Global Affinity Definition

The global link probability is defined as

$$P_{link\_local}(T_i, T_j) = P_{app}(T_i, T_j) P_{gap}(T_i, T_j) P_{temporal}(T_i, T_j). \quad (9)$$

$P_{app}(T_i, T_j)$ is the same as the one defined in Eqt.(8). For the motion link probability $P_{global\_motion}(T_i, T_j)$, a constant velocity assumption was usually made. However, in real complex situations where people have frequent interactions, we cannot expect that every person keeps a constant velocity. Therefore, we use $P_{global\_motion}$ to exclude impossible connections and put more emphasis on the appearance affinity. Specifically, for any two tracklets $T_i$ and $T_j$, if $\left|\mathbf{z}_{i,|T_i|} - \mathbf{z}_{j,1}\right| > (f_{j,1} - f_{i,|T_i|}) d_{max}$, $P_{global\_motion}(T_i, T_j)$ is set to be 0, where $d_{max}$ is the maximum distance that can be traversed by a human object in one time step (1 meter in our experiment). Otherwise, we consider the similarity between the end orientation $\mathbf{o}_{i,|T_i|}$ and the start orientation $\mathbf{o}_{j,1}$ when calculating the motion link probability

$$P_{global\_motion}(T_i, T_j) = \delta + (1-\delta)\max(0, \left\langle \mathbf{o}_{i,|T_i|}, \mathbf{o}_{j,1} \right\rangle), \quad (10)$$

where, $\delta$ controls the importance of $P_{global\_motion}$ in $P_{link\_global}$ and is set to be 0.9.

$P_{gap}(T_i, T_j)$ measures how well the gap between $T_i$ and $T_j$ can be explained and it is defined as

$$P_{gap}(T_i, T_j) = \prod_{k=1}^{f_{j,1} - f_{i,|T_i|} - 1} p_{gap}^{i,j}(k), \quad (11)$$

where $p_{gap}^{i,j}(k)$ calculates how likely the detection at the $k^{th}$ position in the gap is a missed detection. To do this, we first linearly interpolate the real world positions within the gap. Then, for the $k^{th}$ interpolated position, we check if it is occluded by other detections for more than 50%; if it is, this is taken as a missed detection and $p_{gap}^{i,j}$ is set to be the missed detection rate $p_{miss}$, penalizing the missed detection. Otherwise, we check the upper body appearance of the predicted model at this position: If the appearance is similar to both $\mathbf{r}_{i,|T_i|}$ and $\mathbf{r}_{j,1}$, $p_{gap}^{i,j}(k)$ is set to $p_{miss}$ as well; Otherwise, $p_{gap}^{i,j}(k)$ is set to $\eta$ ($<< p_{miss}$), meaning that that gap position cannot be explained by a missed detection and thus it is given a much larger penalty.

### 2.4.2 Global Data Association

Having $P_{link\_global}$ for each tracklet pair, the tracklets association problem is formulated as a MAP problem as proposed in Huang, 2008, which considers track initialization, termination, tracklet association and the probability of tracklets being false alarms. The convergence is guaranteed by reducing the initialization and termination probabilities of each track after each iteration until they reach a predefined lower bound. Figure 5 illustrates an example of the associated tracklets in the global data association step.

The difference in our approach is that we specifically deal with the ambiguous tracklets that may violate the 1st-order Markov chain assumption and thus are likely to introduce identity switches. We consider a tracklet as an ambiguous tracklet when it is linkable to two tracklets at the same end. This type of tracklets usually appears when there are missed detections. In addition, the degenerate tracklets (i.e. tracklets consist of one detection) are also considered ambiguous because they tend to introduce identity switches due to the lack of motion information.

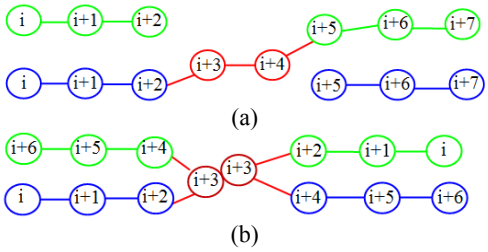



Figure 5: Wrong association (red lines) caused by ambiguous tracklets (red dots). Green dots represents the detections of one human; blue dots represents the detections of another human; red dots represent detections where occlusion happens and ambiguous tracklets are produced. (a) An ambiguous tracklet linkable to two tracklets at each end; (b) two degenerate tracklets that lack motion information.

To approximate 2nd-order Markov chain on the ambiguous tracklets, given the Hungarian association results, we only accept the connection of an ambiguous tracklet at the end with the higher link probability than the other end. The connection of the other end is left for association in the following iterations, when there may be fewer ambiguities (e.g. detection update might have been performed to correct the detections or retrieve the missed detections, or the appearance model may have been updated, or the degenerate tracklet has linked to another tracklet and hence contains motion information).

After the global Hungarian matching, new links may have been established and we can go on performing detection update and local tracklet linking. The iterative process ends when no new links can be found using the global Hungarian association.

### 2.5 Recovery from Identity Switches

Identity switch may exist in the original tracklets, which are usually caused by occlusions where accurate detection is difficult. Within the proposed association framework, as the detection update proceeds, the renewed detection may deviate from the original detection farther and farther away due to the guidance of the reliable temporal information. When the deviation becomes very significant, i.e. the intersection ratio between the updated detection and the original detection is quite small or the appearance affinity between them is not high enough, we doubt that there may be something inconsistent. In this situation, we break up the tracklet at that point and look for possible better association for the resulting two separated tracklets.

## 3 EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of our proposed approach on two public data sets, namely the CAVIAR data set and the PETS 2009 data set, which have been widely used for testing the performance of multi-target tracking works.

In our experiment, parameters not specified manually are learned through 90 ground truth trajectories of a video captured by ourselves where mutual occlusion happens frequently, and these parameters are set exactly the same for both tested data sets.

To determine whether a target is being tracked, the commonly used PASCAL criterion, i.e. the intersection over union greater than 0.5 is adopted for all the experiments.

For quantitative evaluation of the proposed approach, we follow the currently most widely accepted protocol, the CLEAR MOT metrics (Stiefelhagen, 2006): The Multi-Object Tracking Accuracy (MOTA) combines three types of errors – false positives (FP), missed targets (FN), and identity switches (IDs) – and is normalized such that the score of 100% corresponds to no errors (all three error types are weighted equally in our evaluation); The Multi-Object Tracking Precision (MOTP) measures the alignment of the tracker output w.r.t. the ground truth. We also report recall, precision, False alarm per Frame (Fa/F), as well as Mostly Lost (ML), Partially Tracked (PT), and Mostly Tracked (MT) scores, and the number of identity switches (IDs) and fragmentations (Frag) of the produced trajectories compared with ground truth trajectories according to Li, 2009.

Two state-of-the-art tracklet based data association approaches Kuo, 2011 and Yang, 2012 are selected for comparison. In Kuo, 2011, a robust appearance model is learned for each target (PRIMPT), and in Yang, 2012, both appearance models and motion patterns are learned (NLMPRAM). For fair comparison, the detections, ground truth and the evaluation tool are downloaded from the homepage of the first author of Yang, 2012 (http://iris.usc.edu/people/yangbo/downloads.html).

### 3.1 Performance on the CAVIAR Data Set

As the proposed approach requires additional computational time to perform detection update and missed detection recovery, to reduce the run time, for the CAVIAR data set, we sample 1 frame out of every 10 frames from the video sequences for tracking, i.e. the frame rate of the input to the tracking approach is 2.5f/s.

20 sequences of the CAVIAR data set have been evaluated as is done in Kuo, 2011 and Yang, 2012, and Table 1 lists the comparison of the results. It can be seen that our approach outperforms Kuo, 2011 and Yang, 2012 in terms of recall, precision, number of mostly lost tracks and identity switches. However, the number of fragmentations of our approach is higher than both Kuo, 2011 and Yang, 2012. Figure 6 (a) illustrates the tracking result of our approach on CAIVAR data set.

Table 1: Comparison of results on CAVIAR data set.

| Method | Recall | Precision | Fa/F | GT | MT | PT | ML | Frag | IDs |
|---|---|---|---|---|---|---|---|---|---|
| PRIMPT [35] | 88.1 | 96.6 | 0.082 | 143 | 86.0% | 13.3% | 0.7% | 17 | 4 |
| NLMPRAM [36] | 90.2 | 96.1 | 0.095 | 147 | **89.1%** | 10.2% | 0.7% | **11** | 5 |
| Our approach | **91.7** | **97.9** | **0.051** | 147 | 88.4% | 11.6% | **0.0%** | 19 | **3** |

Table 2: Comparison of results on PETS 2009 S2L1 data set.

| Method | Recall | Precision | Fa/F | GT | MT | PT | ML | Frag | IDs |
|---|---|---|---|---|---|---|---|---|---|
| PRIMPT [35] | 89.5% | 99.6% | 0.020 | 19 | 78.9% | 21.1% | 0.0% | 23 | 1 |
| NLMPRAM [36] | 91.8% | 99.0% | 0.053 | 19 | 89.5% | 10.5% | 0.0% | **9** | 0 |
| Our approach | **95.8%** | **99.8%** | **0.013** | 19 | **94.7%** | 5.3% | 0.0% | 21 | 0 |

### 3.2 Performance on the PETS 2009 Data Set

For the PETS 2009 data set, as the sequence was recorded in a low frame rate (7f/s), we did not perform sampling. The comparison result is shown in Table 2. We can see that the recall rate and portion of mostly tracked trajectories have been substantially improved by our approach. However, the number of fragmentations of our approach is still high. Figure 6 (b) shows the tracking result of our approach on PETS 2009 data set.

The high number of fragmentations of our approach is mainly caused by the applied part-based appearance model, for which inaccurate segmentation, which occurs frequently at the spatial temporal locations where occlusion exists, will result in low appearance affinity. In addition, as the appearance model is based on color histogram, it has relatively low discriminability. These two reasons make our approach difficult to deal with some very ambiguous situations. To reduce the possibility of identity switches, a conservative strategy is applied in our approach: if the link probability is low (i.e. the association is likely to introduce identity switches), we choose to discard the association, thus resulting in more fragmentations. We expect that this problem can be much alleviated if more features are used in addition to colors, and discriminative training of appearance models is applied.

### 3.3 Computational Cost Analysis

Our approach is currently realized using MATLAB and implemented on an Intel Corei7 2.93GHz CPU. Most of the computational time is spent on the detection update, which depends on the computational time for each detection update and the total number of detections that need to be updated. For each detection update, usually two times of search for the head position are needed (one for the predicted position and one for the detected position) and the computational time is 1-2 seconds, where the optimal orientation and size are searched within a small neighborhood of the expected orientation $o_0$ and size $s_0$. Then given the head position, the optimal model is selected, where, except for the optimal orientation and size, the leg pose is also searched using the hierarchical model matching as introduced in Wang, 2012. This step usually takes 2-4 seconds. The total number of detection updates depends on the density of the crowd, which is hard to tell, and its upper bound is the total number of human objects in all frames.

The whole association process terminates within 10 iterations for all the tested sequences: for the first several iterations, there are both detection update and local and global tracklet associations; for the remaining iterations, as no detections can be update anymore, only global associations take place.

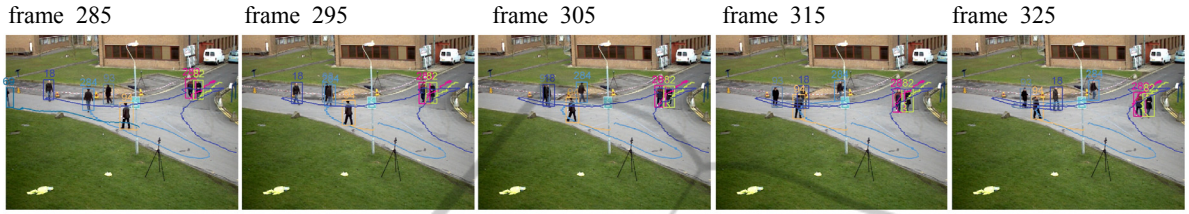
## 4 CONCLUSIONS

In this paper, we propose a hierarchical data association approach that performs detection update using reliable temporal information to improve the accuracy of tracklet quantities. Comparison with two state-of-the-art approaches demonstrates the effectiveness of the proposed approach.

Our future work includes introducing discriminatively trained affinity and appearance

frame 180 frame 220 frame 260 frame 300 frame 340

(a)



frame 285 frame 295 frame 305 frame 315 frame 325

(b)

Figure 6: Illustration of tracking results. (a) CAVIAR data set; (b) PETS 2009 data set.

models into the proposed framework so that fragmentations of the tracking results can be reduced significantly. In addition, our approach can be also improved by exploiting high level scene understanding ability to resolve more ambiguities, e.g. scene occluder detection by either specifically detecting certain commonly seen occluders such as trees and pillars or statistical analysis of the obtained trajectories.

## ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China 61202258, Fundamental Research Funds for the Central Universities of China N100204001, and National Science and Technology Support Program of China 2013BAK02B01-02.

## REFERENCES

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L., 2011. Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33(9), pp. 1820-1833.

Comaniciu, D., Ramesh, V., Meer P., 2003. Kernel-based object tracking. *IEEE Transactions on Pattern Aanlysis and Machine Intelligence,* vol. 25(5) (pp. 564-575.

Huang, C., Wu, B., Nevatia, R., 2008. Robust object tracking by hierarchical association of detection responses. In *Proceedings of European Conference on Computer Vision*, pp. 788-801.

Kuo, C.-H., Huang. C., Nevatia, R., 2010. Multi-target tracking by on-line learned discriminative appearance models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685-692.

Kuo, C.-H., Nevatia, R., 2011. How does person identity recognition help multi-person tracking?. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1217-1224.

Li, Y., Huang, C., and Nevatia, R., 2009. Learning to associate: hybridboosted multi-target tracker for crowded scene. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2953-2960.

Munkres, J., 1957. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics,* vol. 5(1), pp. 32-38.

Perera A., Srinivas C., Hoogs A., Brooksby G., and Hu W., 2006. Multi-object tracking through simultaneous long occlusions and split-merge conditions. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 666-673.

Rasmussen, C., Hager, G. D., 2001. Probabilistic data association methods for tracking complex visual objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(6), pp. 560-576.

Stauffer C., 2003, Estimating tracking sources and sinks. In Proceedings of Computer Vision and Pattern Recognition Workshop, pp. 35.

Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J. S., Mostefa, D., Soundararajan, P., 2006. The CLEAR 2006 valuation. In CLEAR, 2006.

Wang, L., Yung, N., 2012. Three-dimensional model based human detection in crowded scenes. *IEEE Transactions on Intelligent Transportation Systems,* vol. 13(2), pp. 691-703.

Wang, X., Hua, G., Han, T., 2010. Discriminative

Tracking by Metric Learning. In *Proceedings of European Conference on Computer Vision*, pp. 200-214.

Wu, B., Nevatia, R. 2007. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *International Journal of Computer Vision,* vol. 75(2), pp. 247-266.

Yang, B., Nevatia, R., 2012. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1918-1925.