Online Learned Discriminative Part-Based Appearance Models for Multi-human Tracking

Bo Yang and Ram Nevatia

Institute for Robotics and Intelligent Systems,
University of Southern California
Los Angeles, CA 90089, USA
{yangbo,nevatia}@usc.edu

Abstract. We introduce an online learning approach to produce discriminative part-based appearance models (DPAMs) for tracking multiple humans in real scenes by incorporating association based and category free tracking methods. Detection responses are gradually associated into tracklets in multiple levels to produce final tracks. Unlike most previous multi-target tracking approaches which do not explicitly consider occlusions in appearance modeling, we introduce a part based model that explicitly finds unoccluded parts by occlusion reasoning in each frame, so that occluded parts are removed in appearance modeling. Then DPAMs for each tracklet is online learned to distinguish a tracklet with others as well as the background, and is further used in a conservative category free tracking approach to partially overcome the missed detection problem as well as to reduce difficulties in tracklet associations under long gaps. We evaluate our approach on three public data sets, and show significant improvements compared with state-of-art methods.

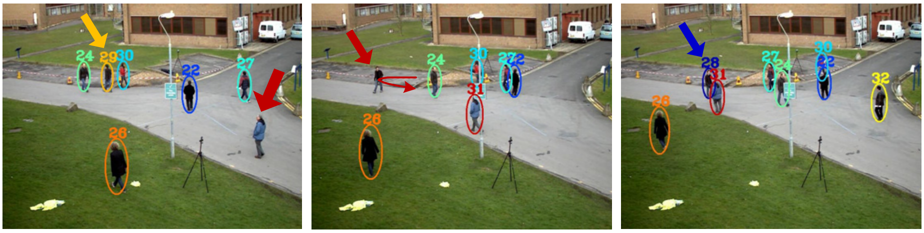
Keywords: multi-human tracking, online learned discriminative models.

1 Introduction

Tracking multiple targets in real scenes remains an important topic in computer vision. Most previous approaches can be classified into Association Based Tracking (ABT) or Category Free Tracking (CFT); ABT is usually a fully automatic process to associate detection responses into tracks, while CFT usually tracks a manual labeled region without requirements of pre-trained detectors. This paper aims at incorporating merits of both ABT and CFT in a unified framework. A key aspect of our approach is online learning discriminative part-based appearance models for robust multi-human tracking.

Association based tracking methods focus on specific kinds of objects, *e.g.*, humans, faces, or vehicles [1–5]; they use a pre-trained detector for the concerned kind of objects to produce detection responses, then associate them into tracklets, *i.e.*, track fragments, and produce final tracks by linking the tracklets in one or multiple steps. The whole process is typically fully automatic. On the contrary, category free tracking methods, sometimes called "visual tracking" in

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 484–498, 2012.



Frame 664

Frame 685

Frame 706

Fig. 1. Limitations of previous association based tracking methods. See text for details.

Table 1. Comparison of association based tracking with category free tracking

	initialization	track solution	motion cue
association based tracking	auto & imperfect	global	available
category free tracking	manual & perfect	individual	unavailable

previous work, continuously track a region based on manual labels in the first frame without requirements of pre-trained detectors [6–8].

We focus on automatically tracking multiple humans in real scenes. As humans may enter or exit scenes frequently, manual initialization is impractical. Therefore, association based tracking is frequently adopted in previous work for this problem [1–4]. In the ABT framework, linking probabilities between tracklet pairs are often defined as

$$P_{link}(T_i \to T_j) = P_a(T_i \to T_j) P_m(T_i \to T_j) P_t(T_i \to T_j) \tag{1}$$

where $P_a(\cdot)$, $P_m(\cdot)$, and $P_t(\cdot)$ denote appearance, motion, and temporal linking probabilities. $P_t(\cdot)$ is often a binary function to avoid temporally overlapped tracklets to be associated, and designs of $P_a(\cdot)$ and $P_m(\cdot)$ play an important role in performance. A global optimal linking solution for all tracklets is often found by a Hungarian algorithm [9, 10] or network flow methods [1, 4].

However, in most previous ABT work, occlusions are not explicitly considered in appearance models for calculating $P_a(\cdot)$, leading to a high likelihood that regions of one person are used to model appearances for another. In addition, ABT only fill gaps between tracklets but does not extend them; hence missed detections at the beginning or end of a tracklet cannot be corrected for associations. An example is shown in Figure 1; the man in blue is missed from frame 664 until frame 685 due to failure of detectors. Moreover, to compute $P_m(\cdot)$ in Equ. 1, it is often assumed that humans move linearly with a stable speed; this assumption would be problematic for long gaps. For example, in Figure 1, person 29 in frame 664 and person 28 in frame 706 are actually the same person but his track is fragmented due to the failure of the detector during the direction change.

One possible solution to overcoming the detection limitation is to use category free tracking methods. However, it is difficult to directly use CFT in association based tracking due to differences in many aspects, as shown in Table 1. First,

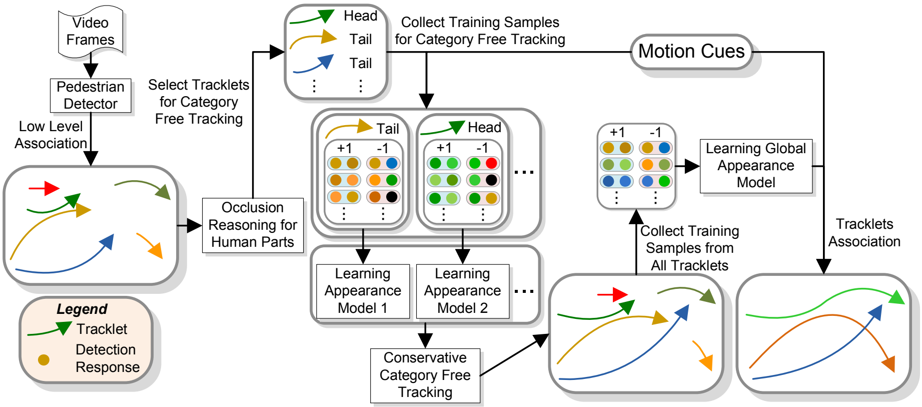

Video Frames
Pedestrian Detector
Low Level Association
Select Tracklets for Category Free Tracking
Occlusion Reasoning for Human Parts
Head
Tail
Tail
Collect Training Samples for Category Free Tracking
Motion Cues
Tail
Head
+1
-1
Learning Appearance Model 1
Learning Appearance Model 2
Conservative Category Free Tracking
Learning Global Appearance Model
Collect Training Samples from All Tracklets
Tracklets Association
Legend
Tracklet
Detection Response

Fig. 2. Tracking framework of our approach. Colors of detection responses come from their tracklets' colors, and black circles denote samples extracted from backgrounds. Head or tail are the earliest or latest parts of a tracklet. Best viewed in color.

CFT starts from a perfect manually labeled region; however, in our problem, automatic initialization is provided by an pre-learned detector, and some detections may be imprecise or false alarms. Second, CFT methods often find best solutions for one or few targets individually; however, in multi-target tracking, a global solution for all targets is more important. Finally, CFT often ignores motion cues, to deal with abrupt motion, while most multi-human tracking problems focus on surveillance videos where people are unlikely to change motion directions and speeds much in a short period, *e.g.*, 4 or 5 frames.

We propose a unified framework that finds global tracking solutions while incorporating the merits of category free tracking with little extra computational cost. We introduce the online learned Discriminative Part-based Appearance Models (DPAMs) to explicitly deal with occlusion problems and detection limitations. The system framework is shown in Figure 2. A human detector is applied to each video frame, and detection responses in neighboring frames are conservatively associated into tracklets. Then based on occlusion reasoning for human parts, we select tracklets that are reliable for CFT. DPAMs for each tracklet are online learned to differentiate the tracklet from backgrounds and other possibly close-by tracklets. A conservative CFT method is introduced to safely track reliable targets without detections, so that missing head or tail parts of tracklets are partially recovered and gaps between tracklets are reduced making linear motion estimations more precise. Finally a global appearance model is learned, and tracklets are associated according to appearance and motion cues.

We emphasize that the category free tracking module in our framework is not proposed to find specific targets from "entry to exit" as most existing CFT methods do, but is used for extrapolating the tracklets and enabling associations to be made more robustly.

The contributions of this paper are:

- A unified framework to utilize merits of both ABT and CFT.
- Part based appearance models to explicitly deal with human inter-occlusions.
- A conservative category free tracking method based on DPAMs.

The rest of the paper is organized as follows: Section 2 discusses related work; building part-based feature sets is described in Section 3; Section 4 introduces the online learning process of DPAMs and the conservative CFT method; experiments are shown in Section 5, followed by conclusion in Section 6.

2 Related Work

Tracking multiple humans has attracted much attention from researchers. To deal with large number of humans, many association based tracking approaches have been proposed [11, 5, 2]. These approaches detect humans in each frame and gradually associate them into tracks based on motion and appearance cues. Appearance models are often pre-defined [9, 4] or online learned [1, 12, 13]. Occlusions are often ignored [13, 2] or modeled as potential nodes in an association graph [4, 1], but have not been used explicitly for appearance modeling, indicating high possibilities that parts used for modeling appearances of a person belong to other individuals. Moreover, performance of ABT is constrained by detection results; missed detections in the beginning or ends of a track cannot be recovered, and tracklets with long gaps between them are difficult to associate according to the commonly used linear motion model [9, 14, 10].

Category free tracking methods do not depend on pre-learned detectors, and appearance models are often based on parts to deal with partial occlusions [6–8]. However, CFT focuses on single or a few objects individually without finding a global solution for all targets, and is difficult to deal with large number of targets due to high complexity. In addition, CFT methods are often sensitive to initialization; once the tracking drifts, it is difficult to recover. If the initialization is imprecise or even a false alarm, the proposed tracks would be problematic.

Note that some previous work also adopts CFT techniques into multi-target tracking problems [9, 3, 15]. However, CFT is often used to generate initial tracklets, which may include many false tracklets or miss some tracklets without carefully tuning the initialization thresholds for each video. However, we use conservative CFT methods to reduce association difficulties and missed detections in an association based framework, but we do not rely it on finding whole tracks. Some previous work also uses parts in tracking [16, 17]; however, only separated persons with few inter-occlusions are considered in [16] while our scenarios are more crowded with frequent occlusions. Parts in [17] are used for detection but not for appearance modeling in tracking as we did.

3 Building Part-Based Feature Sets for Tracklets

In order to improve efficiency, we track in sliding windows one by one instead of processing the whole video at one time. In the following, we will use the term *current sliding window* to refer to the sliding window being processed.

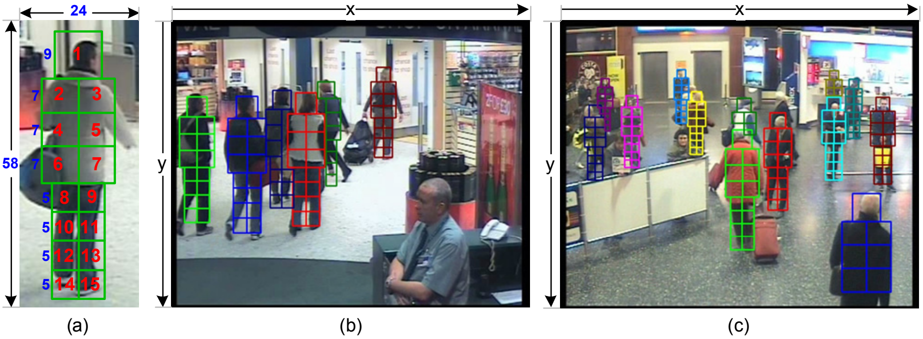


Fig. 3. Illustrations for human parts: (a) parts definition for a normalized 24×58 human detection response; 15 squares are used to build part-based appearance models, their sizes are shown in blue numbers; (b)(c) show automatic occlusion inference results in real scenes; visible parts for each person are labeled with the same color, and occluded parts are not shown and not used for appearance modeling. Best viewed in color.

Given the detection responses in the current sliding window, we adopt the low level association approach as in [13] to generate reliable tracklets from consecutive frames. These tracklets are then associated in multiple steps.

An appearance model for a tracklet T_k includes two parts: a set of features $F_k = \{f_k^1, f_k^2, \ldots, f_k^n\}$ and a set of weights $W = \{w_1, w_2, \ldots, w_n\}$. The features could be color or shape histograms extracted from some regions of responses in the tracklet, and there is a unique F_k for each T_k. The set of weights measures importance of features and is often shared between multiple tracklets. We will use the term *appearance models* to represent the set of weights for clarity. Given W, the appearance similarity between two feature sets F_j and F_k is defined as $\sum_i w_i h_i(f_j^i, f_k^i)$, where $h_i(\cdot)$ is an evaluation function for two features, *e.g.*, Euclidean distances or correlation coefficient between two vectors.

A tracklet T_k is defined as l_k detected or interpolated responses in consecutive frames from time t_k^1 to $t_k^{l_k}$ as $T_k = \{d_k^1, \ldots, d_k^{l_k}\}$, where l_k is the length of T_k. We produce two feature sets F_k^{head} and F_k^{tail}, including features modeling appearances of T_k's head and T_k's tail, *i.e.*, the earliest or latest parts of T_k, respectively. Appearance linking probability from T_k to T_j is evaluated on F_k^{tail} and F_j^{head} using a set of appropriate weights W.

In order to explicitly consider occlusions, features are extracted from parts instead of from whole human responses. As shown in Figure 3(a), we use 15 parts to represent a human so that the parts are not too large to model occlusions and not too small to include meaningful features. Each response d_k^i is normalized into 24×58 and is a union of the 15 parts defined as $d_k^i = \{r_k^i(1), r_k^i(2), \ldots, r_k^i(15)\}$. Each $r_k^i(u)$ is a set of features extracted from part u; for example, $r_k^i(1)$ include the color histogram or the hog feature of part 1. If part u is occluded, all features in $r_k^i(u)$ are invalid, as they may not come from the concerned human. In this section, we aim at building F_k^{head} and F_k^{tail} for each tracklet T_k by explicitly considering human inter-occlusions. Scene occluders, *e.g.*, screens, pillars, are not considered due to difficulty of inferring them.

Algorithm 1. The algorithm of building head feature set for one tracklet

```
Input: tracklet $T_k$.
Initialize $D_k^{head} = \phi$ and $S_k^{head}(j) = \phi \quad \forall j \in \{1, \ldots, 15\}$
For $i = t_k^1, \ldots, t_k^{l_k}$ do:
  – For $j = 1, \ldots, 15$ do:
      • If $|S_k^{head}(j)| < \delta$ and $r_k^i(j)$ is unoccluded, $S_k^{head}(j) = S_k^{head}(j) \cup \{r_k^i(j)\}$;
  – If at least one of $r_k^i(j)$ is unoccluded, $D_k^{head} = D_k^{head} \cup \{d_k^i\}$
  – If $\forall j \in \{1, \ldots, 15\}$, $|S_k^{head}(j)| = \delta$, break;
Compute each feature in $F_k^{head}$ by averaging corresponding features in $S_k^{head}$.
Output: $F_k^{head}$.
```

We assume that all persons stand on the same ground plane, and the camera looks down towards the ground, which is valid for typical surveillance videos. Two scene examples are shown in Figure 3(b)(c). Such assumptions indicate that smaller y-coordinate[1] in a 2D frame corresponds to larger depth from the camera in 3D. Therefore, given detection responses in one frame, we sort them according to their largest y-coordinates indicating their ground plane positions, from largest to smallest. Then we do occlusion reasoning for each person. If more than 30% of a part is occupied by parts of persons who have larger y-coordinates, it is labeled as occluded; otherwise, it is unoccluded. Some examples of automatic detected unoccluded parts are shown in Figure 3(b)(c) in colored squares.

To produce F_k^{head} for tracklet T_k, we decompose the set into 15 subsets as

$$F_k^{head} = \{F_k^{head}(1), F_k^{head}(2), \ldots, F_k^{head}(15)\} \tag{2}$$

Each $F_k^{head}(j)$ is a subset that contains features only extracted from part j. To suppress noise, each feature is taken as the average of valid features from the first δ responses in T_k, where δ is a control parameter and is set to 8 in our experiments. For each part j we introduce a set $S_k^{head}(j)$, which contains multiple $r_k^i(j)$s, *i.e.*, features from part j, from different responses, so that each feature in $F_k^{head}(j)$ is taken as the average value of corresponding features in $S_k^{head}(j)$. Meanwhile, we also maintain a set D_k^{head}, which contains all responses used in building F_k^{head}; this set is used in the learning process for appearance models. For example, D_k^{head} may include the first, second, and eighth responses in tracklet T_k.

Algorithm 1 shows the process of building F_k^{head}, where $|S_k^{head}(j)|$ denotes the number of elements in $S_k^{head}(j)$, and $S_k^{head} = \cup_j S_k^{head}(j)$; the process of building F_k^{tail} is similar. Our algorithm tends to find the earliest δ unoccluded regions for each part. If a part is occluded in early frames, we use later unoccluded ones to represent its appearance, assuming that the occluded parts have similar appearances with those temporally nearest unoccluded ones; this is a widely used assumption in CFT work [18, 8]. If a part j is occluded in the whole tracklet, $F_k^{head}(j)$ would be an empty set, as all features in it are invalid.

[1] See Figure 3(b)(c) for definition of y-coordinates.

4 Online Learning DPAMs for Conservative Category Free Tracking

In this section, we introduce conservative category free tracking methods, so that tracklets are extended from heads or tails to partially overcome the missed detection problem, and long gaps between tracklets are shortened to make linear motion assumptions more reliable.

4.1 Learning of Discriminative Part-Based Appearance Models

As category free tracking in our framework is a conservative process and we do not heavily rely on it to find whole trajectories, we care more about precision than recall. Therefore, we only do CFT for *reliable tracklets*, which satisfy three constraints: 1) they are longer than a threshold ζ, as short tracklets are more possible to be false alarms; 2) the number of non-empty feature set defined in Equ.2 is larger than a threshold β, as appearance models may be unreliable if there are not enough unoccluded parts; 3) it does not reach the boundary of current sliding window. We set $\zeta = 10$ and $\beta = 6$ in our experiments.

If T_k qualifies for CFT from its tail, we online learn discriminative part-base appearance models (DPAMs), so that the learned models well distinguish T_k with other close-by tracklets as well as the background. Far away tracklets are not worth considering as it is difficult to confuse T_k with them.

A linear motion model is often used in association based tracking work. As shown in Figure 4(a), if the tail of tracklet T_k locates at position p_k^{tail}, after time Δt, the human is predicted to be at position $p_k^{tail} + v_k^{tail}\Delta t$, where v_k^{tail} is the velocity of T's tail part. Therefore, we may use the linear motion model to estimate which tracklets are possibly close to T_k in the category free tracking process; we call these tracklets *distracters*.

As the linear motion model is probably invalid over a long period, we do category free tracking only in one second long intervals. If the CFT does not meet termination conditions (detailed later), we do the CFT again for the next one second based on re-learned distracters and appearance models. We expect that the CFT results do not locate too far from the linear estimated positions

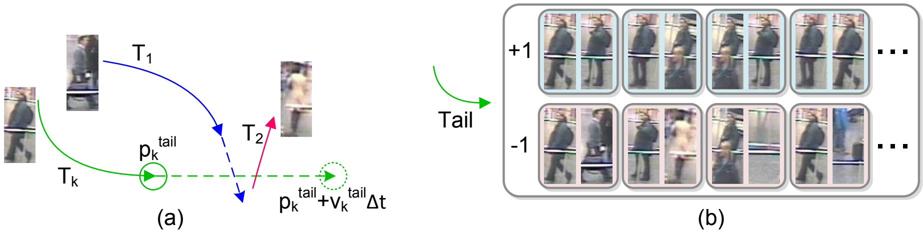


Fig. 4. Online learning of appearance models for category free tracking: (a) estimation of potential distracters for T_k; (b) online collected samples for learning DPAMs

within one second. The estimated distance between T_k and another tracklet T_i at frame t ($t > t_k^{l_k}$) is defined as

$$Dist_t = \begin{cases} \|(p_k^{tail} + v_k^{tail}(t - t_k^{l_k})) - (p_i^{head} - v_i^{head}(t_i^1 - t))\| & \text{if } t < t_i^1 \\ \|(p_k^{tail} + v_k^{tail}(t - t_k^{l_k})) - p_i^t\| & \text{if } t \in [t_i^1, t_i^{l_i}] \\ \|(p_k^{tail} + v_k^{tail}(t - t_k^{l_k})) - (p_i^{tail} + v_i^{tail}(t - t_i^{l_i}))\| & \text{if } t > t_i^{l_i} \end{cases} \quad (3)$$

where $\|\cdot\|$ denotes the Euclidean distance, and p_i^t denotes the response position of tracklet T_i at frame t. A tracklet T_i is a distracter for T_k if it satisfies

$$\exists t \in [t_k^1, t_k^{l_k}] \;\; t_i^1 \le t \le t_i^{l_i} \quad \& \quad \exists t \in (t_k^{l_k}, t_k^{l_k} + \gamma] \;\; Dist_t < \omega h_k^{tail} \quad (4)$$

where γ denotes the frame rate per second, ω is a weight factor, set to 2 in our experiments, and h_k^{tail} denotes the height of T_k's tail response. Equ. 4 indicates that a distracter is a tracklet that has temporal overlap with T_k, so that it belongs to a different human than T_k, and may be close to the future path of T_k, such as T_1 and T_2 in Figure 4(a). Responses in all distracters should have low appearance similarities with responses in T_k; they form a set named as Ψ_k^{tail}.

In addition, to better distinguish T_k from the background, we also collect a set of responses B_k^{tail} from the background in frame $t_k^{l_k}$, defined as

$$B_k^{tail} = \{d_b^l\} \;\; \forall l \in [1, \gamma] \;\; \& \;\; \text{where } d_b^l \text{ is in frame } t_k^{l_k} \;\; \& \;\; p_b^l = p_k^{tail} + v_k^{tail} l \quad (5)$$

The positions of these responses are selected from possible future positions of T_k in the video, but the responses are extracted at the last frame of T_k. Therefore, as long as v_k^{tail} is not zero, these responses do not belong to T_k and should have low appearance similarities with responses in T_k.

Then we build positive and negative training sets $\mathbb{S}^+$ and $\mathbb{S}^-$ for learning DPAMs for T_k's tail, defined as

$$\mathbb{S}^+ = \{x_i = (d_k^{i_1}, d_k^{i_2}), y_i = +1\} \quad \forall d_k^{i_1}, d_k^{i_2} \in D_k^{tail} \quad (6)$$

$$\mathbb{S}^- = \{x_i = (d_k^{i_1}, d_j^{i_2}), y_i = -1\} \quad \forall d_k^{i_1} \in D_k^{tail} \;\; \& \; \forall d_j^{i_2} \in B_k^{tail} \cup \Psi_k^{tail} \quad (7)$$

where D_k^{tail} is the set of responses used in building F_k^{tail} as shown in Algorithm 1. Some visualized examples are shown in Figure 4(b). For a training sample x_i, a control function $v_q(x_i)$ is defined to measure its validity of feature q as

$$v_q(x_i) = \begin{cases} 1 & \text{if } \Phi(f_q) \text{ is not unoccluded in both responses in } x_i \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where $\Phi(f_q)$ denote the part (1 to 15) that a feature f_q belongs to. We adopt the standard RealBoost algorithm to learn discriminative appearance models based on valid features only as shown in Algorithm 2, where $h_q(x)$ denotes the weak classifier, *i.e.*, an evaluation function for two feature qs, *e.g.*, Euclidean distance or correlation coefficient between two vectors, and is normalized to $[-1, 1]$. We adopt features defined in [13], including color, texture, and shape information.

Algorithm 2. The learning algorithm for DPAMs

```
Input: training set $\mathbb{S}^+$ and $\mathbb{S}^-$.
Initialize the weights for samples $w_i = \frac{1}{2|\mathbb{S}^+|}$, if $x_i \in \mathbb{S}^+$; $w_i = \frac{1}{2|\mathbb{S}^-|}$, if $x_i \in \mathbb{S}^-$.
For $t = 1, \ldots, T$ do:
  – For $q = 1, \ldots, n$ do:
      • $r = \sum_i w_i y_i h_q(x_i) v_q(x_i)$
      • $\alpha_q = \frac{1}{2} \ln \frac{1+r}{1-r}$
  – Select $q^* = \underset{q}{\operatorname{argmin}} \sum_i w_i \exp(-\alpha_q y_i h_q(x_i) v_q(x_i))$
  – Set $\alpha_t = \alpha_{q^*}$, $h_t = h_{q^*}$, $v_t = v_{q^*}$
  – Update sample weights $w_i = w_i \exp(-\alpha_t y_i h_t(x_i) v_t(x_i))$, and normalize weights.
Output: $H(x) = \sum_1^T \alpha_t h_t(x)$.
```

4.2 Conservative Category Free Tracking

With online learned DPAMs, we adopt a conservative category free tracking method on each reliable tracklet. The CFT is done one frame by one frame. In each frame, a set of samples are generated around the predicted response by the linear motion model as shown in Figure 5. Sizes and positions of the samples are randomly disturbed around values of the predicted response. A feature set F_s for each sample is extracted to evaluate its similarity to F_k^{tail}. The sample with highest score, named as d^*, is chosen as a potential extension.

However, the potential extension is not taken, if it meets any of the following termination conditions:

- d^* goes beyond the frame boundary or the sliding window boundary.
- The similarity between d^* and F_k^{tail} is smaller than a threshold θ_1.
- The similarity between d^* and any F_i^{head} is larger than a threshold θ_2, where T_i starts at the frame where d^* locates.

In our experiments, we set $\theta_1 = 0.8$ and $\theta_2 = 0.2$. These constraints assures a high similarity between d^* and F_k^{tail} and a low similarity between d^* and any distracters, so that the category free tracking is less likely to drift or go beyond a true association. For example, in Figure 5, T_k and T_2 are actually the same person, and if CFT goes beyond T_2's head, T_k and T_2 cannot be associated, and additional temporal overlapped tracks would be produced for the same person. But if the CFT stops early, we may still successfully associate them.

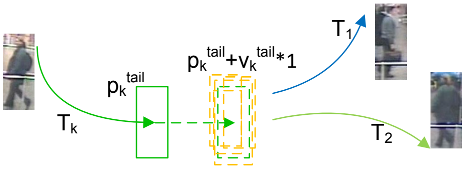


Fig. 5. Illustrations for category free tracking process. The green dashed rectangle is predicted by the linear motion model; the orange dashed ones are samples generated around the predicted position.

Algorithm 3. The CFT algorithm for one tracklet from its tail

```
Input: A reliable tracklet $T_k$ for CFT, and the frame rate per second $\gamma$.
Initialize max frame number for one second tracking $t_{max} = t_k^{l_k} + \gamma$.
Initialize the appearance feature set $F_k^{tail}$ for $T_k$'s tail using Algorithm 1.
While $T_k$ does not meet the boundary of current sliding window do:
  – Find potential distracters in the predicted tracking path of $T_k$ from $t_k^{l_k}$ to $t_{max}$.
  – Collect training samples from $T_k$, its distracters, and the background in the predicted path, and learn DPAMs for the next one second using Algorithm 2.
  – For $t = t_k^{l_k} + 1, \ldots, t_{max}$ do:
      • Generate samples around the predicted position $p_k^{tail} + v_k^{tail} * 1$.
      • Extract the feature set for each sample and evaluate its similarity to $F_k^{tail}$ using the learned DPAMs.
      • Check whether termination conditions are met. If yes, stop CFT; otherwise, add the best response $d^*$ into $T_k$'s tail, and update $p_k^{tail}$ and $v_k^{tail}$
  – Update $t_k^{l_k}$, and set $t_{max} = t_k^{l_k} + \gamma$.
Output: Updated tracklet $T_k$.
```

If d^* is accepted as an extension, we update p_k^{tail} and v_k^{tail} for T_k. However, we do not update F_k^{tail} to avoid drift. This is different with most existing category free tracking approaches, where appearance models are online updated for continuous tracking. In our framework, CFT is only used to partially overcome detection limitations and shorten gaps between tracklets to make linear motion estimation more accurate. If CFT stops early, it is still possible to fill missing gaps by the global association. But errors in CFT may cause failure of associations as discussed above. The whole CFT method is shown in Algorithm 3.

After conservative category free tracking, we learn a global appearance model similar to [13] using only unoccluded parts from all tracklets, and then use the Hungarian algorithm to find global optimal association results.

Figure 6 shows some comparing tracking results by using or disabling the category free tracking module. We can see that person 13 in Figure 6(b) would be fragmented into two tracklets without the CFT due to long time missed detections and non-linear motions. However, the conservative CFT shortens the gaps between the two tracklets so that they become easy to be associated. In addition, person 10 in Figure 6(b) is achieved in frame 330 without having detection responses at that time; person 8 and 15 are similar in Figure 6(d).

5 Experiments

We evaluate our approach on three public data sets: PETS 2009 [19], ETH [20], and Trecvid 2008 [2], which have been commonly used in previous multi-target tracking work. We show quantitative and visualized comparisons with state-of-art methods. We adopt the commonly used evaluation metrics in [10, 2], including recall and precision, showing detection performance after tracking; false alarms per frame (FAF); number of trajectories in the ground truth (GT); mostly tracked (MT), mostly lost (ML), and partially tracked (PT), denoting ratios of tracks with successful tracked parts for more than 80%, less than 20%, and others; fragments (Frag), the number of times that a ground truth track is interupted; id switches (IDS), the number of times that a produced track changes its matched ground truth track.

The three data sets have different resolutions, densities, and average motion speeds. However, we use the same parameter settings on all, and performances all improve compared with state-of-art methods. This indicates low sensitivity of our approach on parameters.

As detection performance would influence tracking results, for fair comparisons, we use the same detection results as in [13, 10, 2] on three data sets, which are provided by authors of [2, 10]. No scene occluders are manually assigned.

5.1 Results on PETS 2009 Data Set

We use the same PETS 2009 video clip as used in [21] for fair comparison. The target density is not high in this set. However people are frequently occluded by each other or scene occluders, and may change motion directions frequently.

The quantitative comparison results are shown in Table 2. We modified the ground truth annotations from [21], and fully occluded people that appear later are still labeled as the same person. We can see that by only using part based appearance models, the MT is improved by more than 10%, and fragments are reduced by 43% compared with up-to-date results in [10]. By using category free tracking, we further improve MT by 5%, and reduce fragments by 85%. This indicates our part based models are effective for modeling humans' appearances, and the conservative CFT method shortens gaps between tracklets so that fragmented tracks can be associated based on the linear motion model. Some visualized examples are shown in Figure 6(b).

5.2 Results on ETH Data Set

The ETH data set [20] is captured by a pair of cameras on a moving stroller in busy street scenes. The heights of human may change significantly from 40 pixels to more than 400 pixels. The cameras shift frequently making the linear motion model less reliable, and there are frequent full inter-occlusions due to the low camera angle. We use the same ETH sequences as in [10]. Only data captured by the left camera are used in our experiment.

The quantitative results are shown in Table 3. We can see that using the part models only, MT is improved for about 8%; using category free tracking improves MT significantly for an additional 12%. In ETH data, partial and full occlusions are frequent; therefore, the part based models help build more precise appearance models. The category free tracking method recovers many missed detections, especially for humans who appear for only short periods, *e.g.*, less than 40 frames, and therefore improves MT significantly.

Some tracking examples are shown in Figure 6(d) and Figure 7(a). Person 10 in Figure 7(a) is detected from frame 151 until frame 158, but we start to track her from frame 140 until frame 162. After frame 162, our conservative CFT stops tracking due to large appearance changes caused by the shadows.

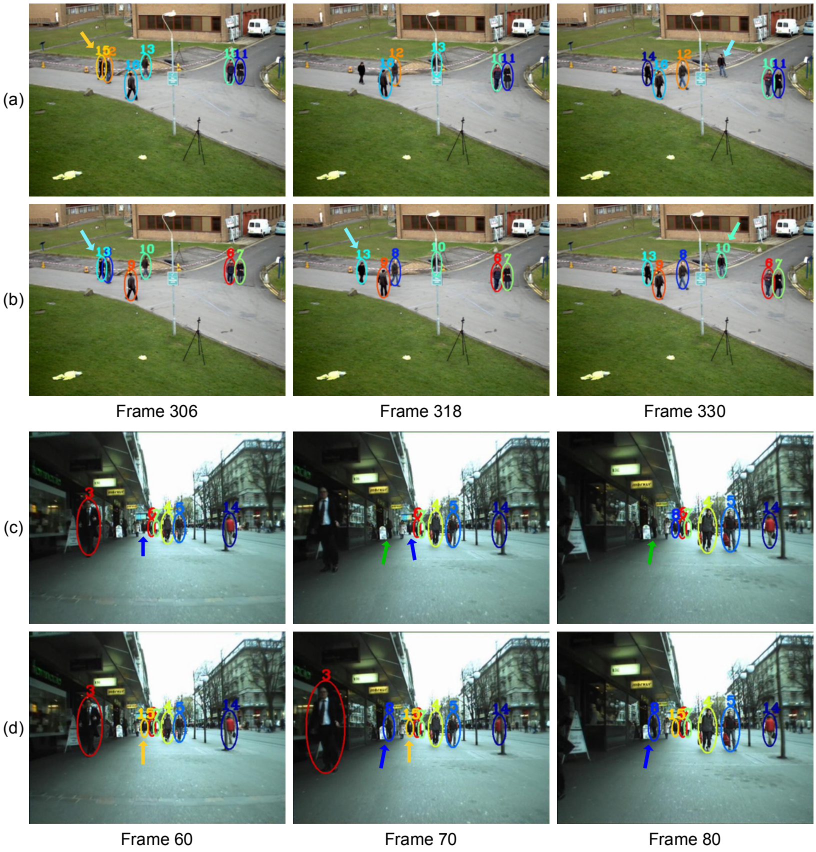


Fig. 6. Comparisons of tracking results with or without category free tracking: (a)(c) show results without CFT, (b)(d) show results on the same sequences with CFT

Table 2. Comparison of results on PETS 2009 dataset. The PRIMPT results are provided by authors of [10]. Our ground truth is more strict than that in [21].

Method	**Recall**	**Precision**	**FAF**	**GT**	**MT**	**PT**	**ML**	**Frag**	**IDS**
Energy Minimization [21]	-	-	-	23	82.6%	17.4%	0.0%	21	15
PRIMPT [10]	89.5%	99.6%	0.020	19	78.9%	21.1%	0.0%	23	1
Part Model Only	92.8%	95.4%	0.259	19	89.5%	10.5%	0.0%	13	0
Part Model + CFT	97.8%	94.8%	0.306	19	94.7%	5.3%	0.0%	2	0

5.3 Results on Trecvid 2008 Data Set

Trecvid 2008 is a very difficult data set, which contains 9 video clips with 5000 frames for each. The videos are captured in a busy airport; the density is very

Table 3. Comparison of tracking results on ETH dataset. The human detection results are the same as used in [10], and are provided by courtesy of authors of [10].

Method	**Recall**	**Precision**	**FAF**	**GT**	**MT**	**PT**	**ML**	**Frag**	**IDS**
PRIMPT [10]	76.8%	86.6%	0.891	124	58.4%	33.6%	8.0%	23	11
Part Model Only	77.5%	90.9%	0.595	124	66.1%	25.0%	8.9%	21	12
Part Model + CFT	81.0%	87.8%	0.861	124	78.2%	12.9%	8.9%	19	11

Table 4. Comparison of tracking results on Trecvid 2008 dataset. The human detection results are the same as used in [2, 13, 10], and are provided by authors of [2].

Method	**Recall**	**Precision**	**FAF**	**GT**	**MT**	**PT**	**ML**	**Frag**	**IDS**
Offline CRF Tracking [2]	79.2%	85.8%	0.996	919	78.2%	16.9%	4.9%	319	253
OLDAMs [13]	80.4%	86.1%	0.992	919	76.1%	19.3%	4.6%	322	224
PRIMPT [10]	79.2%	86.8%	0.920	919	77.0%	17.7%	5.2%	283	171
Part Model Only	78.7%	88.2%	0.807	919	73.0%	20.9%	6.1%	253	149
Part Model + CFT	79.2%	87.2%	0.895	919	75.5%	18.6%	5.9%	247	145

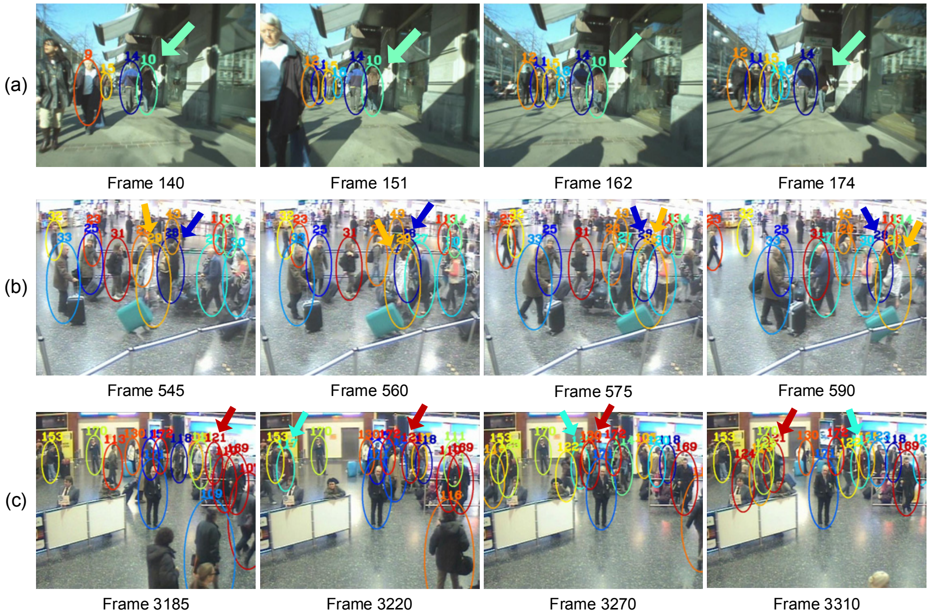
high, and people occlude each other frequently. Quantitative comparison results are shown in Table 4. Compared with [10], using only part based models reduces fragments and id switches by about 11% and 13% respectively; using CFT additionally reduces fragment and ID switches by about 2% and 3%. In Trecvid 2008 data sets, most improvements come from part based models and less from the category free tracking method, which is quite different with results on PETS 2009 and ETH. This is because Trecvid 2008 is a very crowded data set; in the CFT process, there are often many distracters for each tracklet. Therefore the CFT often stops early because the probability is quite high that at least one of many distracters would have similar appearances with the concerned person.

Figure 7(b)(c) show some tracking examples. We see that person 28 & 29 in Figure 7(b) and person 121 in Figure 7(c) are under heavy occlusions due to high densities of humans; however, our approach is able to find correct discriminative part-based appearance models, and therefore tracks these persons successfully.

5.4 Computational Speed

The processing speed is highly related with the number of humans in videos. We implement our approach using C++ on a PC with 3.0GHz CPU and 8GB memory. The average speeds are 22fps, 10fps, and 6fps on PETS 2009, ETH, and Trecvid 2008 respectively. Compared with [10], which reported 7fps on Trecvid 2008, our approach does not impose much extra computational cost[2]. The major extra cost is from the feature extractions for all samples in the CFT module, which could be constrained by setting the number of evaluated samples.

[2] Detection time is not included in both our reported speed and that in [10].



Frame 140 Frame 151 Frame 162 Frame 174

Frame 545 Frame 560 Frame 575 Frame 590

Frame 3185 Frame 3220 Frame 3270 Frame 3310

Fig. 7. Tracking results of our approach on ETH and Trecvid data sets

6 Conclusion

We introduced online learned discriminative part-based appearance models for multi-human tracking by incorporating merits of association based tracking and category free tracking. Part-based models are able to exclude occluded regions in appearance modeling, and a conservative category free tracking can partially overcome limitations of detection performance as well as reduce gaps between tracklets in the association process. Experiments on three public data sets show significant improvement with little extra computational cost.

Acknowledgments. Research was sponsored, in part, by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-10-2-0063. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: ICCV (2011)
2. Yang, B., Huang, C., Nevatia, R.: Learning affinities and dependencies for multi-target tracking using a crf model. In: CVPR (2011)

3. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: CVPR (2011)
4. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR (2011)
5. Song, B., Jeng, T.-Y., Staudt, E., Roy-Chowdhury, A.K.: A Stochastic Graph Evolution Framework for Robust Multi-target Tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 605–619. Springer, Heidelberg (2010)
6. Wang, S., Lu, H., Yang, F., Yang, M.H.: Superpixel tracking. In: ICCV (2011)
7. Liu, B., Huang, J., Yang, L., Kulikowsk, C.: Robust tracking using local sparse appearance model and k-selection. In: CVPR (2011)
8. Grabner, H., Matas, J., Gool, L.V., Cattin, P.: Tracking the invisible: Learning where the object might be. In: CVPR (2010)
9. Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: CVPR (2009)
10. Kuo, C.H., Nevatia, R.: How does person identity recognition help multi-person tracking? In: CVPR (2011)
11. Pellegrini, S., Ess, A., Schindler, K., Gool, L.V.: You'll neverwalk alone: Modeling social behavior for multi-target tracking. In: ICCV (2009)
12. Stalder, S., Grabner, H., Van Gool, L.: Cascaded Confidence Filtering for Improved Tracking-by-Detection. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 369–382. Springer, Heidelberg (2010)
13. Kuo, C.H., Huang, C., Nevatia, R.: Multi-target tracking by on-line learned discriminative appearance models. In: CVPR (2010)
14. Song, X., Shao, X., Zhao, H., Cui, J., Shibasaki, R., Zha, H.: An online approach: Learning-semantic-scene-by-tracking and tracking-by-learning-semantic-scene. In: CVPR (2010)
15. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 1820–1833 (2011)
16. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking people by learning their appearance. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 65–81 (2007)
17. Wu, B., Nevatia, R.: Detection and segmentation of multiple, partially occluded objects by grouping, merging, assigning part detection responses. International Journal of Computer Vision 82, 185–204 (2009)
18. Kalal, Z., Matas, J., Mikolajczyk, K.: P-n learning: Bootstrapping binary classifiers by structural constraints. In: CVPR (2010)
19. Pets 2009 dataset (2009), http://www.cvg.rdg.ac.uk/PETS2009
20. Ess, A., Bastian Leibe, K.S., van Gool, L.: Robust multiperson tracking from a mobile platform. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 1831–1846 (2009)
21. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: CVPR (2011)