Coupling Multiple Alignments and Re-ranking for Low-latency Online Multi-target Tracking

Yingkun Xu*, Lei Qin*, Qingming Huang*‡

*Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
‡University of Chinese Academy of Sciences, Beijing, China
{yingkun.xu, lei.qin, qingming.huang}@vipl.ict.ac.cn

Abstract. Previous works for multi-target tracking employ two strategies: global optimization and online state estimation. In time-critical applications, the former methods have long temporal latency, and the latter can't recover from erroneous association or drifting. In this paper, we combine these two strategies, and propose a new low-latency online tracking approach. Unlike previous multi-hypotheses methods, which are always suffered from combinational explosion, our approach keeps the candidate associations using multiple alignments only in ambiguous cases. The novel features based on previous multi-frame associations are designed for re-ranking of the multiple linkages. The experimental results illustrate the advantage and robustness of these features based on prediction of previously generated tracks, and their discrimination to find optimal ones. Comparison with five state-of-the-art methods proves that our proposed method is competitive to global optimal ones and is superior to other online tracking algorithms.

1 Introduction

Visual multi-target tracking is a very important topic in computer vision. Applications based on visual multi-target tracking can be roughly classified into two categories. One is related with offline analysis after events take place. Mining similar actions and searching specific activities are examples of this category. The other aims to react to the online time-critical scenarios, such as finding abnormal events immediately or predicting dangerous accidents, and so on. Our work aims to promote the performance of multi-target tracking for the second category applications.

Various tracking approaches [1-8] are proposed to handle the tracking problems based on associations of detection responses. Considering the relationships between neighbor detection responses as being linked or not, the detection association can be modeled as network flow problems [1], k-shortest paths optimization [2], or conditional random fields (CRFs) with different constrains [3, 4]. There are large feasible solution spaces for these models, and global optimizations are employed to achieve promising performance. In order to employ these global

optimization algorithms for online time-critical applications, the temporal sliding windows mechanism is utilized to extend these methods. However, they are suffered from the cost of long latency by the straightforward extension.

In order to implement online multi-target tracking for quick reaction in instantaneous tasks, the frame-by-frame associations using greedy or bipartite matching approaches are employed with well-designed affinity metrics [5, 6]. By comparison with global optimization, these approaches are more likely to produce some ambiguous associations at current instants. Because it is more distinguishable considering the linkages of future frames, one possible solution is using the information of next few frames to decide the better association at the present moment. Thus, the current tracking results will be deferred for some frames. If this latency is small enough, for example less than 0.5 seconds, such as the latency has nearly no effect for online applications, we still can deem the algorithm as online ones. Differing from frame-by-frame online tracking, we call these online methods with small-deferring as low-latency online tracking algorithms. The classic low-latency online tracking algorithms are multi-hypothesis tracking method (MHT) [7] and joint probabilistic data association filter (JPDAF) [8]. However, these two methods are suffered from two problems. The first is combinational explosion when the space of observation increases. The second is the final determination of association is coupled tightly with the local features which cause the ambiguities, thus the erroneous linkages can seldom be corrected. We need global features beyond the local affinities to find a better choice. In our work, we propose a new multi-association based online multi-tracking method. Instead of only using Gaussian kernel similarity of position in MHT, we combine different appearance and motional features in structured output Ranking-SVM framework, and generate multiple hypotheses considering different alignments between tracks and detections. The final association is determined by the multi-frames features considering the previous long temporal multi-frame assignments as illustrated in Figure 1.

The insight of our approach tries to hybridize the frame-to-frame local affinity and multi-frame associations into one low-latency online tracking framework. We utilize weighted multiple features for frame-to-frame matching as in [6]. Through using multiple alignments, we can explore wider association space, and have more probability to cover correct linkages. We select the best solution using high-order multiple frame features which are more informative and discriminative than frame-to-frame affinity. Those long-term association features are widely used in the CRF model [3, 4]. However, it is time-consuming and even intractable to train and to infer the best association status using arbitrarily complex features in CRFs. By comparison, re-ranking the possible associations using these multi-frame association features is more efficient and concise. Thus, the low-latency online solution can be designed using the strategy of local multi-frame associations and global re-ranking. In our method, the weights for combing the multi-frame association features are learned offline, and our approach is based on a hybrid strategy of the online tracking and offline learning.

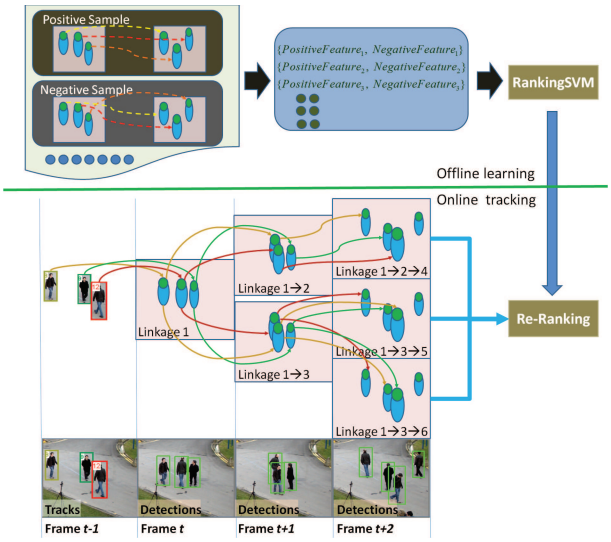

Fig. 1. The framework of our proposed approach. We construct the association pairs from ground truth as training samples to learn RankingSVM model, and use the model to re-rank the current candidate associations when they can't be decided definitely. The 3 frame associations are composed by 3 step linkages as linked number list (such as linkage $1 \rightarrow 2 \rightarrow 4$) in the figure. The selected association is considered as the best association, and the its first linkage is the t-frame tracking result.

The main contributions of this paper lie in three points. (I) We develop a novel low-latency online multi-target tracking framework using re-ranking strategy. Compared with global optimization, which directly infer the most probable posterior within the CRF or MRF framework, we can use more complex high-order features. (II) We propose a new method to generate online multiple-hypotheses without bringing in the problem of combinational explosion. These multiple hypotheses consider the multiple alignments between tracks and detections, and are only necessary in the case of ambiguous linkage. (III) The discriminative features are proposed for multi-frame association re-ranking. Experimental results prove they are effective for re-ranking of the candidate associations based on previous tracking results.

2 Related Works

The core motivation of our paper is how to integrate the global association into the online tracking framework. Our strategy lies from the spirit of deferring the decision when the online matching is not easy to obtain. We instantiate this strategy in the online multi-target tracking with structured learning and re-ranking algorithm. The related works are outlined in following aspects.

Generating a small set of feasible solutions, and finding the best one after accumulation of evidences is a common strategy in the field of visual computing.

It is multi-stage cascading procedure in essence. The face detection algorithm [9] illustrates its advantages in performance and speed. This strategy is applied for measurement appraisal in the MHT [7] and JPDAF [8]. These approaches depend on the two stages of finding the candidate K-best solutions and choosing the optimal one with largest sum of log-probabilities of leaf-branches after a temporal delay. The cascading algorithms have also been utilized within particle filter framework [10, 11]. Different from these algorithms, we combine the re-ranking algorithm to promote the association results of the online structured learning.

Combination of offline global optimization and online updating is utilized in some previous works [12, 13]. It has been proved to be effective to improve the tracking performance using the selected features which are trained with large scale dataset. In [12], a most discriminative feature pool is learned beforehand, and they serve as the candidate features for each gallery track segment. In [13], a deep stacked denoising auto-encoder is employed to learn the robust features from an image dataset, and these features are updated in online tracking using additive sigmoid classifiers. By contrast, our proposed approach can be considered as a tradeoff between the global optimization and online multi-target state updating.

There have been some works related with getting M best solutions in probabilistic model [14, 15], and deterministic graphic optimization problems [16, 17]. In our work, we use different alignments to get at most M best solutions within the framework of structured output learning. Among the candidate M best solutions, we utilize the re-ranking algorithm to score and select the potential best one. Re-ranking algorithms are transferred from natural language processing [18] and information retrieval [19] to the problems of visual tracking [20, 21]. In [20], a CRF model is constructed to represent the possible connections between detection responses. RankBoost algorithm is employed to train the model with sampling association pairs. In [21], the weakly supervised ranking algorithm is proposed to learn the weights of appearance features. The graph Laplacian is used to regularize the smoothness of similarities between samples. Different from above methods, we employ the high-order and complex features from multi-frames associations to preferably assess the correctness of the candidate solutions, and appraisal the best one.

3 Online Multi-target Tracking with Multiple Alignments between Tracks and Detections

Learning multiple features and combining them is an important topic to enhance the robustness of online multi-target tracking. Here, we employ the structured output SVM learning as in [6] to learn the weights for combining multiple features.

Denoting the j-th detection response in frame t as $s_j^t = (b_j^t, o_{j,1}^t, o_{j,2}^t, \ldots, o_{j,K}^t)$, where b_j^t is the bounding box and $o_{j,k}^t$ is its k-th feature, the set of detection responses in frame t is $S^t = \{s_j^t\}_{j=1}^n$. Further, we can define the i-th candidate

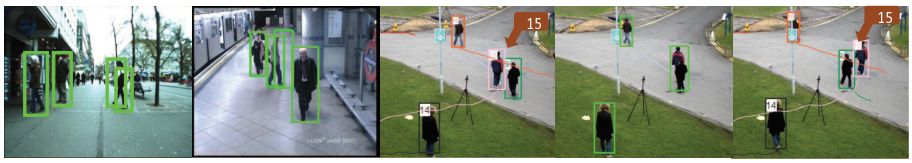

Fig. 2. (a)(b) Detection responses are imprecise using two type detectors [23, 24]. (c)(d)(e) Object 15 has occluded template features. In order to decide the association of object 15 to the detections in (d), the association of next frame in (e) has more distinguishable features.

track till frame $t-1$ as the detection response list $r_i^{t-1} = \{s_i^{t-l}, s_i^{t-l+1}, \ldots, s_i^{t-1}\}$ from the start frame $t-l$ to frame $t-1$. Thus, the set of candidate tracks for frame t is $R^{t-1} = \{r_i^{t-1}\}_{i=1}^{m}$. Based on the K dimension features defined for each detection response, we can calculate the K-dimensional affinity between r_i^{t-1} and s_j^t denoted as $a_{i,j}^t = \phi(r_i^{t-1}, s_j^t)$, and combine the K-dimensional affinity vector using weight w into one affinity scalar.

In the problem of online multi-target tracking using detection response association, we need to link the candidate tracks R^{t-1} and detection responses S^t for each frame t. Let a binary vector set $\mathcal{Y}_t = \left\{y_t | y_t = [y_{1,1}^t, \ldots, y_{m,1}^t, y_{1,2}^t, \ldots, y_{m,n}^t]^T\right\}$ represent the linking result candidates, there should be the constraints $\sum_i y_{i,j} \leq 1$ and $\sum_j y_{i,j} \leq 1$ because one candidate track is matched with one detection response at most. To solve the problem of how to obtain the online association y^t. The optimal bipartite matching method can be employed to obtain y_t by $y_t = \arg\max_{y \in \mathcal{Y}_t} \langle w, y^T \Phi(R^{t-1}, S^t) \rangle$, where the feature matching matrix $\Phi(R^{t-1}, S^t) = [a_{1,1}^t, \ldots, a_{m,1}^t, a_{1,2}^t, \ldots, a_{m,n}^t]^T$. To obtain the optimized weight vector $\tilde{w}$, we utilize the structured output SVM algorithm as in [6]:

$$\begin{aligned} \tilde{w} = \arg\min_{w} \quad & \frac{1}{2}\|w\|^2 + \frac{C_1}{N}\sum \xi^t \\ s.t. \ \max_{y \in \mathcal{Y}_t} & \Delta(y, y_t) - \langle w, y_t^T \Phi(R^{t-1}, S^t) - y^T \Phi(R^{t-1}, S^t) \rangle \leq \xi^t \\ & \forall \xi^t \geq 0, t = 1...N \end{aligned} \tag{1}$$

This problem has efficient cutting-plane solution [27] using N collected samples: $\{y_t, \Phi(R^{t-1}, S^t)\}_{t=1}^{N}$.

The above formulations present an efficient solution for multi-target tracking by offline learning. Same as in [6], we define the loss function $\Delta(y, y_t) \equiv y^T(1-y_t)$ and 42-dimensional features for each detection response. However, there are still some problems needed to be discussed in detail. One of the most important issues is how to define the K-dimensional affinity $a_{i,j}^t = \phi(r_i^{t-1}, s_j^t)$ so that the matching between s_j^t and r_i^t is aligned. In order to design meaningful and effective affinity vector, we should keep the features in tracks and in detections consistent as far as possible. Normally, it is difficult for detectors to obtain complete aligned results. It is also hard to design consistent tracking features using detection

pools along tracks, which acts as the templates to match with detections. For example, in Figure 2, the detectors [23, 24] obtain some imprecise results as illustrated in (a) (b). To match with detection results in the next frame as in (c) (d), the tracks may utilize their last detected features as their templates because these detections are mostly close to their next matched detections. However, this strategy is not suitable for the case of partial occlusion, in which the features of the last detections may be generated from the occluders rather than from targets themselves as in (d). These misaligned features between tracks and detections make the affinity unreliable for online tracking.

From above discussion, the problems of inconsistent features between tracks and detections can be considered as different cases of misalignments. From the view of tracking, the features of tracks may suffer from unreliability due to pose changing and occlusions, which will cause the feature templates to be ambiguous. These ambiguities can be considered as progressive misalignment. From the view of detections, the bounding boxes of detection response may be skew to one side because of irregular shapes or non-max suppression operations. These skew bounding boxes can be considered as detection misalignment. These two kinds of misalignments are the main problems of how to design the appropriate affinity functions between tracks and detections. Our paper mainly discuss how to obtain the better online association using re-ranking considering these misalignments.

4 Re-ranking of Multi-online Associations with SSVM Learning

4.1 Re-ranking of Multiple Online Associations using Multi-alignments

To handle the feature misalignment between tracks and detections, a straightforward approach is to infer the best one among the candidate alignments. However, it is nearly intractable to find the optimal alignment because it is hard to predict which trajectories are occluded and whether the detections are irregular. Therefore, we keep the different alignments to cover the optimal one as possible as we could. In our approach, the kept alignments of tracks have two types: the average features and the latest features of their most recent detections. The kept alignments of detections are based on four corners of the bounding box. Therefore, there are at most eight cases for the association in one frame, although in the most frames only one or two of them are different and kept.

To find the best one of the multiple alignments at the frame t, we foresee associations in next Δt frames based on previous ΔT tracking results, and we denote the temporal range from $t - \Delta T$ to $t + \Delta t$ as $[t - \Delta T : t + \Delta t]$. Because the accumulative features are more discriminative than those from single one frame, we accumulate evidences of multiple frame association in these $\Delta T + \Delta t$ frames. Then, we employ the re-ranking algorithm with the accumulative features to select the best candidate multi-frame associations. To simplify the detailed explanation of our re-ranking algorithm. We extend the notations in section 3 and list their descriptions in Table 1.

Table 1. Notations for Re-Ranking of Multi-Frame Associations

Symbols	Description
$[t_a : t_b]$	The discrete set from t_a to t_b. If t_a and t_b is frame number, then the discrete set represent one frame range from t_a to t_b.
$R_{[t-\Delta T:t+\Delta t]}$	The trajectory set which are overlapped with the frame range from frame $t-\Delta T$ to frame $t+\Delta t$.
$S_{[t-\Delta T:t+\Delta t]}$	The detection set which are detected within the frame range from frame $t-\Delta T$ to frame $t+\Delta t$.
$y_{[t:t+\Delta t]}$	The association between trajectory set $R_{[t-\Delta T:t+\Delta t]}$ and detection set $S_{[t:t+\Delta t]}$ within the frame range from t to $t+\Delta t$.
$r_k^{[t-\Delta T:t+\Delta t]}$	The k-th trajectory in the trajectory set $R_{[t-\Delta T:t+\Delta t]}$ which has detection set within the frame range from $t-\Delta T$ to $t+\Delta t$.
t_s^k, t_e^k	The first frame number and the last frame number of the trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ within the frame range from $t-\Delta T$ to $t+\Delta t$.
$s_k^{t_p}$	The detection response at the frame t_p for the trajectory $r_k^{[t-\Delta T:t+\Delta t]}$.
$B(r_k^{[t-\Delta T:t]})$	One B-spline curve fitting for the k-trajectory $r_k^{[t-\Delta T:t]}$.
$\theta_k^{t_p}$	The motion angle of the k-trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ at frame t_p.
$P_k^{t_p}$	The location point of detection bounding box of the k-trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ at frame t_p.
$W_k^{t_p}$	The width of detection bounding box of the k-trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ at frame t_p.
$v\theta(\Delta t_\tau)_k^{t_p}$	The angle velocity of the k-trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ at frame t_p.
$vP(\Delta t_\tau)_k^{t_p}$	The linear velocity of the k-trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ at frame t_p.

The function to score the multi-frame associations based on above multiple alignments is expressed as:

$$f_{[t-\Delta T:t+\Delta t]}(y_{[t:t+\Delta t]}) = \alpha^T \Psi(R_{[t-\Delta T:t+\Delta t]}, S_{[t-\Delta T:t+\Delta t]}, y_{[t:t+\Delta t]}) \quad (2)$$

where α is the weight of re-ranking features Ψ, $S_{[t-\Delta T:t+\Delta t]}$ is the set of detections in temporal range $[t-\Delta T : t+\Delta t]$, and $R_{[t-\Delta T:t+\Delta t]}$ are trajectories in $[t-\Delta T : t+\Delta t]$ by appending the foresee associations $y_{[t:t+\Delta t]}$. The best association of y_t can be obtained by $y_t = \arg\max_{y\in\tilde{\mathcal{Y}}_t} f_{[t-\Delta T:t+\Delta t]}(y_{[t:t+\Delta t]})$, in which $\tilde{\mathcal{Y}}_t$ is the candidate association set obtained from section 3 considering multi-alignments. Next subsection will discuss how to design the re-ranking features Ψ.

4.2 Re-ranking Features for Multi-frame Associations

The key to find optimal choice of candidate associations is to design appropriate features $\Psi(R_{[t-\Delta T:t+\Delta t]}, S_{[t-\Delta T:t+\Delta t]}, y_{[t:t+\Delta t]})$ for these associations. Intuitively, the foreseeing short-term associations $y_{[t:t+\Delta t]}$ in frames $[t : t+\Delta t]$ are expected to be consistent with the previous long-term associated trajectories backward to frames. To achieve this point, we design the features in three

aspects. First, we expect the observed detection responses in each trajectory between frames $[t : t + \Delta t]$ have consistent appearance and motional trends as its previous part in frames in frames $[t - \Delta T : t]$. Second, the relationship between every two trajectories should have potential consistency with the scenario, while keeping the exclusions between each other. Third, the statistical attributes of the associations should have similar distribution as those in the training dataset. We discuss in detail for these features in following three aspects.

In the first aspect, we consider the features to keep intra-consistency for trajectories from the points of appearance, shape and motion. Given the trajectory set $R_{[t-\Delta T:t+\Delta t]} = \{r_k^{[t-\Delta T:t+\Delta t]}\}_{k=1}^{M}$, we split the frame length ΔT to N_a parts $[t - \Delta T_i : t - \Delta T_{i+1}]$, $i \in [1..N_a]$, using same log-length frame intervals. We denote the k-th track as the detection list $\{s_k^{t_s^k}, ..., s_k^{t_e^k}\}$. For each interlaced frame $t_j \in [t : 2 : t + \Delta t]$, the appearance intra-consistency feature can be expressed as $\Psi_{i,j}^1$:

$$\Psi_{i,j}^1 = \frac{1}{M}\sum_{k=1}^{M} I(t_s^k < t - \Delta T_i) \max_{t_p \in [t-\Delta T_i : t-\Delta T_{i+1}]} Aff(s_k^{t_p}, s_k^{t_j}) \tag{3}$$

where $I(.)$is indicator function, and $Aff(s_k^{t_p}, s_k^{t_j})$ is the similarity between two detections $s_k^{t_p}$, $s_k^{t_j}$ using Bhattacharyya coefficient. Using three types of appearance information (HSV, LBP, and RGB color), we can obtain 45 dimension features for appearance similarities based on 3 interlaced short-term frames and 64 frames of previously tracking results ($N_a = 5$).

To express the intra-consistency of the shape for trajectories, we expect the smoothness of fitting curves is good as possible as those in ground truth. Thus, fitting the detection points and computing the errors of detection points in the future Δt frames is very important clues. To allow motional changes, we fit the curve for detections within different length frames, with the length as half of the previous length. For example, looking backward 64 frames of previously tracking results, there are 4 groups of curves $B_i(r_k^{[t-\Delta T:t]})$, where the i-th group of curves is obtained from temporal range $[t - 2^i\Delta\tau : t]$, $i \in [1..4]$. Thus, for the i-th group of curves, the smoothness for all tracks can be:

$$\Psi_i^2 = \frac{1}{M}\sum_{k=1}^{M} I(t_s^k < t - 2^{i-1}\Delta\tau)\frac{1}{\Delta t}\sum_{t_j \in [t:t+\Delta t]} \exp(-\lambda_1 \left\| P_k^{t_j} - B_i(r_k^{[t-\Delta T:t]}) \right\|_2) \tag{4}$$

where $\|P - B\|_2$ is the Euclidean distance from the center point P to the curve B.

To illustrate the intra-consistency of motional trend for the trajectories, we consider the motional direction and velocity respectively. Given the directions and locations of the trajectory $r_k^{[t-\Delta T:t+\Delta t]}$ as $\{\theta_k^{t_s^k}, \ldots, \theta_k^{t_e^k}\}$ and $\{P_k^{t_s^k}, \ldots, P_k^{t_e^k}\}$, we can compute their angular velocity and linear velocity at each moment as $\{v\theta(\Delta t_\tau)_k^{t_i^k} = (\theta_k^{t_i^k} - \theta_k^{t_i^k+\Delta t_\tau})/\Delta t_\tau\}$ and $\{vP(\Delta t_\tau)_k^{t_i^k} = (P_k^{t_i^k} - P_k^{t_i^k+\Delta t_\tau})/\Delta t_\tau\}$ by consideration of different temporal interval Δt_τ. It is helpful to tolerate the misalignments of detections using different Δt_τ because the small perturbations

exist in the detection responses. Assuming the angular velocities obey von Mises distribution as in [25] and the linear velocities obey Gaussian distribution, the motion consistencies based on temporal difference Δt_τ are expressed as:

$$\Psi^3_{\Delta t_\tau} = \frac{1}{M}\sum_{k=1}^{M}\frac{1}{\Delta t}\sum_{t_j\in[t:t+\Delta t]}\frac{1}{2\pi I_0(\lambda_2)}\exp(\lambda_2\cos(v\theta(\Delta t_\tau)_k^{t_j})) \tag{5}$$

$$\Psi^4_{\Delta t_\tau} = \frac{1}{M}\sum_{k=1}^{M}\frac{1}{\Delta t}\sum_{t_j\in[t:t+\Delta t]}\exp(-\frac{(vP(\Delta t_\tau)_k^{t_j} - vP(\Delta t_\tau)_k^{t_j-1})^2}{2\sigma_k^2(vP(\Delta t_\tau)_k)}) \tag{6}$$

where I_0 is modified Bessel function of order 0. Using 3 different values for Δt_τ, e.g. (1, 2, 4), we can get 6 dimension features.

In the second aspect, we expect to obtain the features between every two trajectories. These features should reflect the information of mutual exclusion among the trajectories, as well as coincident with context. For example, people always walk along the limited paths since there are not too many roads in one scenario. Thus, some persons will pass along similar paths when there are lots of people walking through. We design the features from the aspects of trajectory shapes. To obtain the features between the trajectory shapes, we compute the chamfer distance between every two trajectories. We keep its 5-bins histogram after normalization by the width of bounding box $\{W_k^{t_s^k},\ldots,W_k^{t_e^k}\}$. The value of i-bin is:

$$\Psi_i^5 = \frac{1}{M(M-1)}\sum_{u\neq v}^{M} I(L(i) < Chd(r_u^{[t-\Delta T:t+\Delta t]}, r_v^{[t-\Delta T:t+\Delta t]}) < H(i)) \tag{7}$$

$$\text{where}: \quad \mathrm{Chd}(\mathrm{r_u},\mathrm{r_v}) = \frac{1}{|\mathrm{r_u}|}\sum_{\mathrm{t_j}\in[\mathrm{t_t^u}:\mathrm{t_e^u}]}\min_{\mathrm{t_l}\in[\mathrm{t_t^v}:\mathrm{t_e^v}]}(\left\|\mathrm{P_u^{t_j}} - \mathrm{P_v^{t_l}}\right\|/\mathrm{W_u^{t_j}})$$

where $L(i)$ and $H(i)$ are the low and high boundary for bin i. We set the values as $(0, 0.5, 1, 2, 4)$ and $(0.5, 1, 2, 4, +\infty)$ for 5 bins respectively.

In the third aspect, extra statistic features of the trajectory group in the temporal range $[t-\Delta T : t+\Delta t]$ are calculated according to the ground truth. We expect the features obtained from the test associated trajectories are matched to these statistic values as much as possible. We utilize three statistic features. The first is the average occluded length. Inspired by the work [4], we use Cauchy-Lorentz distribution to model the consecutive occluded frames of each trajectory. The average weighted value is:

$$\Psi^6 = \frac{1}{M}\sum_{k=1}^{M}\prod_{\Delta j\in gaps(r_k^{[t-\Delta T:t]})}\frac{\lambda_3}{\left|r_k^{\Delta j}\right|^2+\lambda_3^2} \tag{8}$$

where $gaps(r)$ is the occluded segments of track r, and $\left|r_k^{\Delta j}\right|$ is the consecutive frame length for occluded segment Δj.

The second statistic information is the number of started trajectory and the number of terminated trajectory between the temporal range $[t : t+\Delta t]$.

We assume they are modeled by exponential distribution. Thus, we obtain the features related with these numbers as:

$$\Psi^7 = \exp(-\lambda_4 \frac{1}{M} \sum_{k=1}^{M} I(t <= t_s^k < t + \Delta t)) \tag{9}$$

$$\Psi^8 = \exp(-\lambda_5 \frac{1}{M} \sum_{k=1}^{M} I(t <= t_e^k < t + \Delta t)) \tag{10}$$

The third statistic information is the length of trajectory. We hope the trajectory extends as long as possible. Thus, too short trajectories are unexpected because they are always obtained by abnormal linkages. We calculate this attribution as following feature:

$$\Psi^9 = \exp(-\lambda_6 \frac{1}{M} \sum_{k=1}^{M} \frac{1}{t_e^k - t_s^k}) \tag{11}$$

Above 9 types of features are appended to appraisal whether the generated trajectories are better or not. The parameters $(\lambda_1, \lambda_2, \ldots, \lambda_6)$ are estimated by the ground truth. Using the length of $\Delta t = 5$ and $\Delta T = 64$, there are 64-dimension features by considering all above three aspects' features. By experiments, they are suitable for discrimination between positive associated tracks and negative ones.

4.3 Ranking SVM Learning of Multi-online Associations

Because our goal is to appraise multi-frame associations obtained using multiple alignments, we need to prioritize them using one score function $f_{[t-\Delta T:t+\Delta t]}(.)$. This is a problem of re-ranking learning for structured output as discussed previously [26]. Given one pair of association results $y^i_{[t:t+\Delta t]}$ and $y^j_{[t:t+\Delta t]}$ such that the former has higher priority than the latter, denoting as $y^i_{[t:t+\Delta t]} \succ y^j_{[t:t+\Delta t]}$, we hope their values using score function have relationship:

$$y^i_{[t:t+\Delta t]} \succ y^j_{[t:t+\Delta t]} \Leftrightarrow f_{[t-\Delta T:t+\Delta t]}(y^i_{[t:t+\Delta t]}) > f_{[t-\Delta T:t+\Delta t]}(y^j_{[t:t+\Delta t]}) \tag{12}$$

Considering the linear weighted formulation defined in equation (2) and features defined in above section, we need to obtain the optimal weight vector α. This re-ranking for structured output can be solved efficiently using cutting-plane algorithm [27]. We employ the scaling slack form to learn the weight vector:

$$\alpha = \arg\min_{\alpha} \frac{1}{2} \|\alpha\|^2 + \mathrm{C}_2 \sum_{\mathrm{i} \in [1:|Q|]} \zeta^i \tag{13}$$

$s.t.$

$$\alpha^T (\Psi(R_{[t_i-\Delta T:t_i+\Delta t]}, S_{[t_i-\Delta T:t_i+\Delta t]}, y^\circ_{[t:t+\Delta t]}) - \Psi(R_{[t_i-\Delta T:t_i+\Delta t]}, S_{[t_i-\Delta T:t_i+\Delta t]}, y_{[t:t+\Delta t]})) \geq 1 - \frac{\zeta^i}{L(y^\circ_{[t:t+\Delta t]}, y_{[t:t+\Delta t]})}$$

$$\forall \zeta^i \geq 0, \quad \forall y_{[t:t+\Delta t]} \neq y^\circ_{[t:t+\Delta t]} \quad \wedge \quad y_{[t:t+\Delta t]} \in Y_{[t:t+\Delta t]}$$

In our method, we extract the training sample pairs set Q from the ground truth. Each sample in this set is composed by the detections $S_{[t_i-\Delta T:t_i+\Delta t]}$ with at most length $\Delta T + \Delta t$, their association results $y^{\circ}_{[t:t+\Delta t]}$ by changing some linkages in ground truth as introduced in experiments, and the associated tracks $R_{[t_i:t_i+\Delta t]}$. The loss $L(y^{\circ}_{[t:t+\Delta t]}, y_{[t:t+\Delta t]})$ between two associations $y^{\circ}_{[t:t+\Delta t]}$ and $y_{[t:t+\Delta t]}$ are defined as the Hamming loss of the $y_{[t:t+\Delta t]}$ when $y^{\circ}_{[t:t+\Delta t]}$ is same with ground truth, or difference between the Hamming losses of $y_{[t:t+\Delta t]}$ and $y^{\circ}_{[t:t+\Delta t]}$. We use efficient one-slack algorithm [27] to train the weight vector α.

5 Experiments

To evaluate the performance of our proposed method, we utilize three public datasets: PETS09-S2-L1, ETHMS, and TUD. The sequences in these datasets contain different visual conditions, such as static camera and moving camera, partial occlusion and full occlusion, pose variation, and illumination changing, etc. More importantly, the detection results and tracking ground truth of these datasets are opened to public.[1] Thus, we can compare with other methods fairly.

We set the hyper-parameters C_1 and C_2 as 10.0 and 100.0, and estimate other parameters related with specific distributions by training dataset. In experiments, we train the re-ranking of multiple candidate associations using multi-frame features (3)-(11) in offline process. Then, we compare with five state-of-the-art multi-target tracking algorithms with public evaluation metrics.

In the offline training process, we learn the weights α for association features Ψ. We employ the ranking SVM algorithm. First, we need to construct the training dataset Q which is composed by multi-frame association pairs $\{(y^{q1}_{[t:t+\Delta t]}, y^{q2}_{[t:t+\Delta t]})\}_{q=1}^{|Q|}$. In each frame range $[t : t + \Delta t]$, we extract three different associations $y^{GT}_{[t:t+\Delta t]}$, $y^{A}_{[t:t+\Delta t]}$ and $y^{B}_{[t:t+\Delta t]}$, where the first association $y^{GT}_{[t:t+\Delta t]}$ is same as ground truth, and latter two associations $y^{A}_{[t:t+\Delta t]}$ and $y^{B}_{[t:t+\Delta t]}$ are obtained by three different transformation operations to $y^{GT}_{[t:t+\Delta t]}$: switching some detection segments between two tracks, drifting some detections for some tracks, or adding some faked tracks into the association. By using these operations, the constructed associations have such relationship $y^{GT}_{[t:t+\Delta t]} \succ y^{A}_{[t:t+\Delta t]} \succ y^{B}_{[t:t+\Delta t]}$ that we can construct three different pair samples $(y^{GT}_{[t:t+\Delta t]}, y^{A}_{[t:t+\Delta t]})$, $(y^{GT}_{[t:t+\Delta t]}, y^{B}_{[t:t+\Delta t]})$ and $(y^{A}_{[t:t+\Delta t]}, y^{B}_{[t:t+\Delta t]})$. The size of training dataset triples the number of frame ranges which we can separate training sequences into. Thus, we construct 1404, 1095 and 114 pair samples from ETHMS, PETS09 and TUD respectively, and verify the tracking results by cross-validation between them.

Our method aims to combine the re-ranking multi-candidate associations into online tracking strategy. As discuss above, we need the cross-validate strategy to evaluate our method using different training set for re-ranking learning. We call our method as re-ranking based low-latency online multi-target tracking (ReRankingLMT). Specifically, the methods based on training datasets, ETHMS, PETS09 and TUD, are named ReRankingLMT-E, ReRankingLMT-P

[1] http://iris.usc.edu/people/yangbo/downloads.html.

and ReRankingLMT-T respectively. To evaluate the performance of our proposed method using this combining strategy, we need to compare with the methods using online strategy and offline optimization. Using the common ground truth and detection response input, we compare with five state-of-the-art methods. The first two methods are recognition based tracking (PIRMPT) [12] and structured output SVM based method (SSVMMOT) [6]. Similar with our methods, these two algorithms combine the offline learned features to online tracking process. The last three methods are energy based algorithm (EnergyMIN) [28], the method considering exclusion between detections and trajectories (ExcTracking) [4] and the online CRF based method (OnlineCRF) [3]. These three methods construct different CRFs with different constraints, and global optimizations are executed to finding the best association results. It is convincible for our proposed method to compare with these five methods in aspects of both feature learning and long-term association optimization.

To evaluate the quantitative performance, we employ the VACE metrics [3]. These metrics are mainly composed by detection recall (RECALL), detection precision (PREC), the percentage of the mostly tracked objects (MT), the percentage of the partial tracked objects (PT), the percentage of the mostly lost objects (ML), the number of trajectories' interruption by tracking (Frag), and the number of real identities' changes for tracked trajectory (IDS). Because ML is redundant with MT and PT, we omit this item. These metrics can be calculated using public tools [3]. Moreover, we use the harmonic mean (F) of RECALL and PRECISION to reflect the overall metric.

Table 2 gives the results of comparison. By comparison with the similar online algorithms, PIRMPT and SSVMMOT, which utilized offline learned features in the online tracking manner. Our method exceeds them both in recall and integrity of the trajectories. Most true detections are linked to our tracked results and trajectories are generated more completely. By comparison with global optimization methods, EnergyMIN, ExcTracking and OnlineCRF, our approach shows competitive results. In the scenario of the static camera, such as in PETS09, our approach even outperforms these global optimizing algorithms. There are less identity switches and false-alarming tracking fragments. As discussed above, these attribute to the stable distinguishability in ambiguous associations. Besides, different from these global optimizing methods which all induce long-time delay when used for online tracking applications, our method only has low latency less than 10 frames.

Figure 3 shows some tracking examples. The first row illustrates some results from the static camera as in PETS09-S2-L1 sequence. There are interactions and short term partial occlusions, such as those of person 11 and person 9 in the 480-th and 510-th frames. Our method appears robust for these partial occlusions, even in the long-time occlusion case for person 3 in the 65-th frame. The second and third rows illustrate the examples when camera is parallel with view field in TUD and ETHMS sequences. There are many inaccurate detection results in TUD sequence, and person 3 and person 8 pass behind others causing full occlusions. Our method obtains correct tracking trajectories for them, although

Table 2. Comparison With State-of-the-art Methods

Datasets	METHOD	RECALL	PREC	F	MT	PT	Frag	IDS
PETS-2009-S2-L1	PIRMPT[12]	89.5%	99.6%	0.94	78.9%	21.1%	23	1
	SSVMMOT[6]	97.2%	93.7%	0.95	94.7%	5.3%	19	4
	OnlineCRF[3]	93.0%	95.3%	0.94	89.5%	10.5%	13	0
	ExcTracking[4]	—	—	—	94.7%	5.3%	15	22
	EnergyMIN[28]	92.4%	98.4%	0.95	91.3%	4.3%	6	11
	ReRankingLMT-E	98.9%	97.5%	0.98	94.7%	5.3%	3	2
	ReRankingLMT-T	98.9%	97.7%	0.98	94.7%	5.3%	4	2
TUD Stadt-mitte	PIRMPT[12]	81.0%	99.5%	0.89	60.0%	30.0%	0	1
	SSVMMOT[6]	80.0%	96.7%	0.88	80.0%	20.0%	11	0
	OnlineCRF[3]	87.0%	96.7%	0.92	70.0%	30.0%	1	0
	ExcTracking[4]	—	—	—	40.0%	60.0%	13	15
	EnergyMIN[28]	84.7%	86.7%	0.86	77.8%	22.2%	3	4
	ReRankingLMT-E	89.0%	98.6%	0.94	80.0%	20.0%	3	3
	ReRankingLMT-P	89.5%	98.2%	0.94	80.0%	20.0%	4	3
ETHMS	PIRMPT[12]	76.8%	86.6%	0.81	58.4%	33.6%	23	11
	SSVMMOT[6]	78.4%	84.1%	0.81	62.7%	29.6%	72	5
	OnlineCRF[3]	79.0%	90.4%	0.84	68.0%	24.8%	19	11
	ExcTracking[4]	77.3%	87.2%	0.82	66.4%	25.4%	69	57
	ReRankingLMT-P	79.7%	86.4%	0.83	66.0%	24.5%	37	33
	ReRankingLMT-T	78.9%	86.9%	0.83	62.8%	27.7%	38	27

several ID switches are induced such as for track 5 in the 76-th frame. By contrast with TUD, the sequence of ETHMS is taken from moving camera, and there are lots of full occlusions when the persons pass by. Our method can handle them in most cases like in frame 910 to frame 950, where two persons 77 and 78 are full occluded and recovered after person 73 leaves the view of camera.

We implement our method with runtime of about 2.2FPS using the non-optimized code of Matlab 2013 at the platform of INTEL i7-3632QM and 2.2GHz CPU. This runtime is slow mostly due to the calculation of multi-frame features for multi-candidate associations. However, it can be promoted in the future since most of these calculations are redundant. Another issue is how to handle the problem of occlusion. Because we mainly focus on how to find better candidate association in this paper, we adopt the method in [6] to handle the short-term occlusion problem for each association. Besides, we expect the occlusion gaps are reasonable due to the feature (8). One of the advantages of our method is we can integrate our re-ranking process into the other online tracking frameworks, because our method is not tightly coupled with them, and promote the final performance.

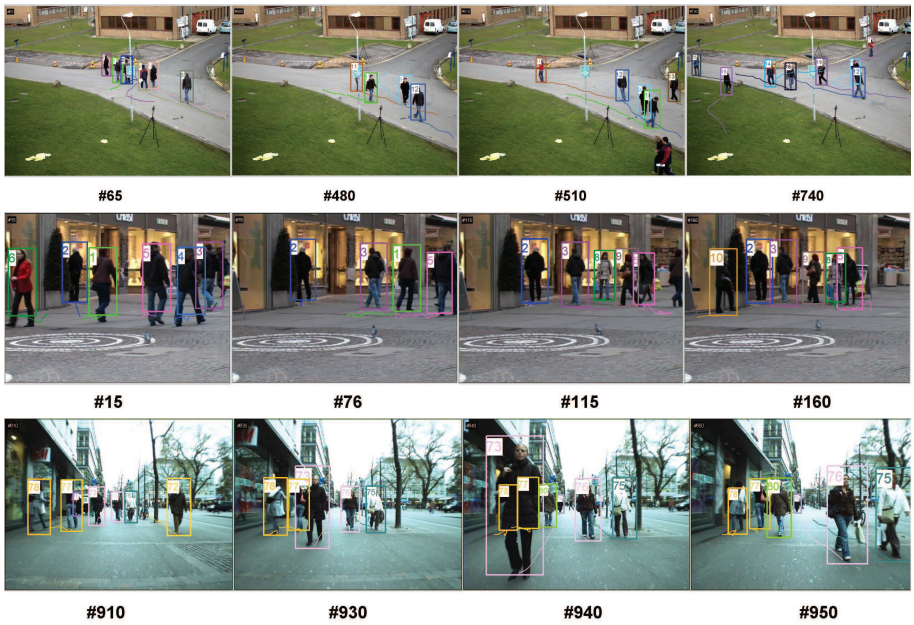

Fig. 3. Tracking examples for (a) PETS09, (b) TUD and (c) ETHMS bahnnof.

6 Conclusions

In this paper, we propose an online tracking framework combining the global optimization of previous long-term associations. We implement this by re-ranking of multiple short-term associations generated by multiple alignments. The features considering both the trajectory specific self-similarities and mutual relationships between tracks are designed to choose the optimal associations. The experimental results show the discrimination of these weighted features by offline training. By comparison with five state-of-the-art algorithms in three public datasets using common evaluation tools, our method outperforms the other online tracking algorithms and is competitive with global optimal ones. In the future work, we can try to promote the performance by adding more related features, or integrate the possible alignments into an efficient online optimization algorithm by treating them as latent variables, and thus avoiding the re-ranking process.

Acknowledgement. This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400, in part by National Natural Science Foundation of China: 61025011, 61133003, 61332016, 61390510.

References

1. Zhang, L., Li, Y., Nevatia, R.: Global Data Association for Multi-Object Tracking Using Network Flows. In: Proc. CVPR (2008)
2. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. PAMI (2011), vol. 33(9), pp. 1806-1819.
3. Yang, B., Nevatia, R.: Multi-Target Tracking by Online Learning a CRF Model of Appearance and Motion Patterns. IJCV (2013)
4. Milan, A., Schindler, K., Roth, S.: Detection- and Trajectory-Level Exclusion in Multiple Object Tracking. In: Proc. CVPR (2013)
5. Yan, X., Wu, X., Kakadiaris, I. A., Shah, S. K.: To Track or To Detect? An Ensemble Framework for Optimal Selection. In: ECCV (2012)
6. Kim, S., Kwak, S., Feyereisl, J., Han, B.: Online Multi-Target Tracking by Large Margin Structured Learning. In: ACCV (2012)
7. Reid, D. B.: An algorithm for tracking multiple targets. IEEE Trans. on Automatic Control (1979)
8. Bar-Shalom, Y., Formann, T. E.: Tracking and Data Association. Academic Press, San Diego, CA (1988)
9. Viola, P., Jones, M. J.: Robust real-time face detection. IJCV (2004)
10. 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 life spans. PAMI, 30, 1728-1740 (2008)
11. Stalder, S., Grabner, H., Gool, L. V.: Cascaded confidence filtering for improved tracking-by-detection. In: ECCV (2010)
12. Kuo, C. H., Nevatia, R.: How does person identity recognition help multi-person tracking? In: Proc. CVPR (2011)
13. Wang, N., Yeung, D. Y.: Learning a Deep Compact Image Representation for Visual Tracking. In: NIPS (2013)
14. Fromer, M., Globerson, A.: An LP View of the M-best MAP problem. In: NIPS (2009)
15. Batra, D., Yadollahpour, P., Guzman-Rivera, A., Shakhnarovich, G.: Diverse M-best solutions in Markov random fields. In: ECCV (2012)
16. Guerriero, F., Musmanno, R., Lacagnina, V., Pecorella, A.: A class of label-correcting methods for the K shortest paths problem. Operations Research, 49, 423-429 (2001)
17. Murty, K. G.: An algorithm for ranking all the assignments in order of increasing cost. Operating Research, 16, 682-687 (1968)
18. Collins, M., Koo, T.: Discriminative reranking for natural language parsing. Computational Linguistics, 31, 25-70 (2005)
19. Liu, T. Y.: Learning to rank for information retrieval. Foundations and Trends in Information Retrieval, 3, 225-331 (2009)
20. Yang, B., Huang, C., Nevatia, R.: Learning Affinities and Dependencies for Multi-Target Tracking using a CRF Model. In: Proc. CVPR (2011)
21. Bai, Y., Tang, M.: Robust tracking via weakly supervised ranking svm. In: Proc. CVPR (2012)
22. Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., Singer, Y.: Online passive-aggressive algorithms. JMLR (2006)
23. Huang, C., Nevatia, R.: High performance object detection by collaborative learning of joint ranking of granules features. In: Proc. CVPR (2010)

24. Felzenszwalb, P. F., Girshick, R. B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. PAMI, 32, 1627-1645 (2010)
25. Song, B., Jeng, T. Y., Staudt, E., Roy-Chowdhury, A. K.: A stochastic graph evolution framework for robust multi-target tracking. In: ECCV (2010)
26. Joachims, T.: Training Linear SVMs in Linear Time. In: KDD (2006)
27. Joachims, T., Finley, T., Yu, C. N. J.: Cutting-plane training of structural SVMs. Machine Learning, 77, 27-59 (2009)
28. Milan A., Roth S., Schindler K.: Continuous Energy Minimization for Multi-Target Tracking. PAMI, 36, 58-72 (2014)