Tracklet Association with Online Target-Specific Metric Learning

Bing Wang, Gang Wang, Kap Luk Chan, Li Wang
School of Electrical and Electronic Engineering, Nanyang Technological University
50 Nanyang Avenue, 639798, Singapore
{wang0775,wanggang,eklchan,wa0002li}@ntu.edu.sg

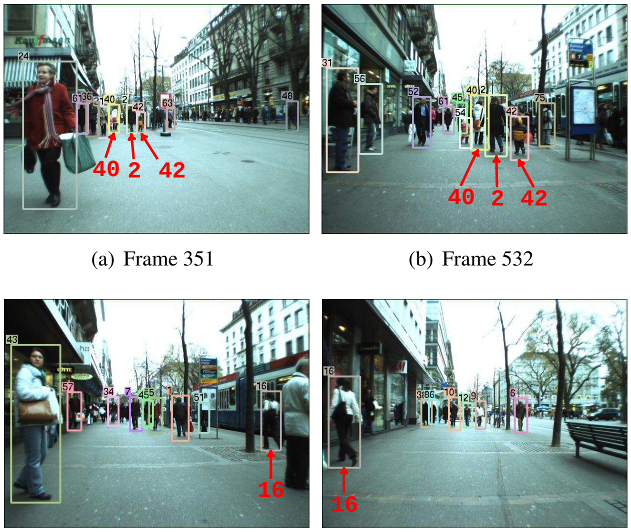
Abstract

This paper presents a novel introduction of online target-specific metric learning in track fragment (tracklet) association by network flow optimization for long-term multi-person tracking. Different from other network flow formulation, each node in our network represents a tracklet, and each edge represents the likelihood of neighboring tracklets belonging to the same trajectory as measured by our proposed affinity score. In our method, target-specific similarity metrics are learned, which give rise to the appearance-based models used in the tracklet affinity estimation. Trajectory-based tracklets are refined by using the learned metrics to account for appearance consistency and to identify reliable tracklets. The metrics are then relearned using reliable tracklets for computing tracklet affinity scores. Long-term trajectories are then obtained through network flow optimization. Occlusions and missed detections are handled by a trajectory completion step. Our method is effective for long-term tracking even when the targets are spatially close or completely occluded by others. We validate our proposed framework on several public datasets and show that it outperforms several state of art methods.

1. Introduction

In this paper, we address the challenging problems in long-term tracking of multiple persons in a complex scene captured by a single, uncalibrated camera. The challenging problems are due to many sources of uncertainty, such as clutter, serious occlusions, targets interactions, and camera motion.

Recently, significant progress has been reported in human detection [6, 8, 7, 10, 21, 22, 23], and this promotes the popular tracking paradigm: detect-then-track [11, 13, 24, 28, 3, 19, 4, 2, 5]. The main idea is that a human detector is run on each frame to detect targets of interest, and then detection responses are linked across multi-frames to obtain target trajectories. In [28, 3, 19, 5], the authors formulate the multi-frame, multi-target data association as a network flow problem. Zhang et al. [28] use a push-relabel method [9] to solve the min-cost flow problem. Berclaz et al. [3] and Pirsiavash et al. [19] propose to use more efficient successive shortest path algorithms, which can provide roughly the same globally optimal tracking results with less running time. In a more recent paper, Butt et al. [5] incorporate higher-order track smoothness constraints, such as constant velocity, for multi-target tracking. However, due to the limitation of the deployed appearance cues, such methods usually cannot deal with long-term tracking to obtain a complete trajectory of a target. This is because frequent prolonged occlusions and target interactions will result in track fragments of a trajectory. If we make full use of the information from the whole sequence (previous, current, and subsequent frames), trajectory can be recovered and track-



(a) Frame 351 (b) Frame 532

(c) Frame 673 (d) Frame 877

Figure 1. Frames from ”BAHNHOF” sequence of ETH dataset with target identities labeled by our method. The ID labels 2, 40, 42 in the top row and ID label 16 in the bottom row remain unchanged after many occlusions and interactions over more than 180 frame intervals.

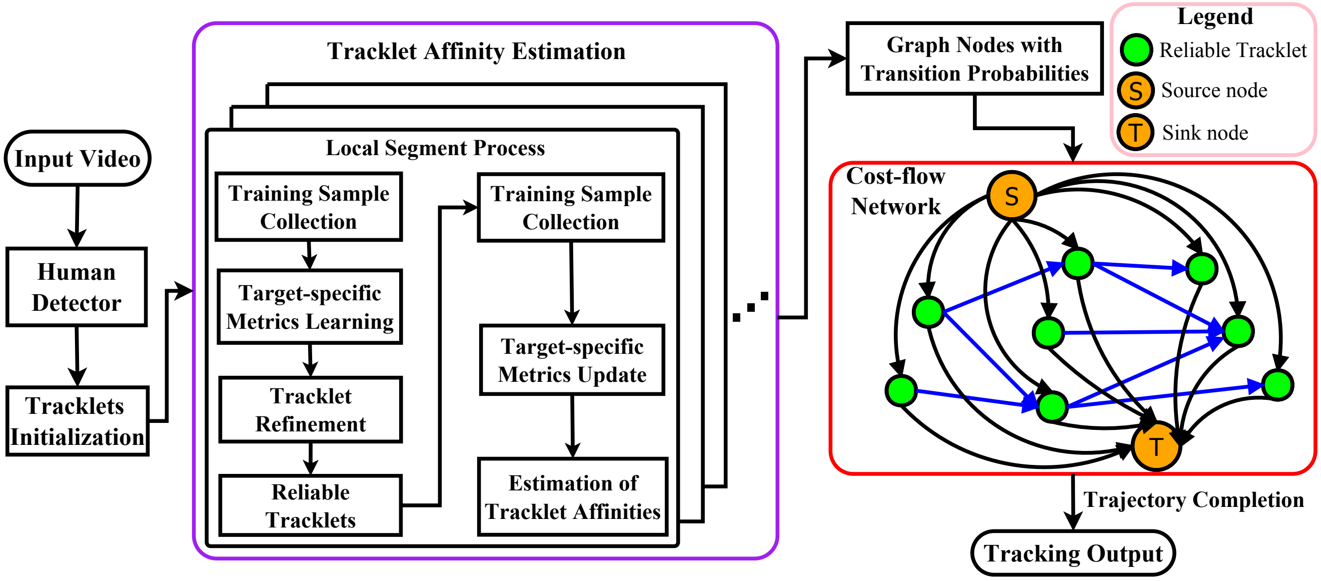

Input Video
Human Detector
Tracklets Initialization
Tracklet Affinity Estimation
Local Segment Process
Training Sample Collection
Target-specific Metrics Learning
Tracklet Refinement
Reliable Tracklets
Training Sample Collection
Target-specific Metrics Update
Estimation of Tracklet Affinities
Graph Nodes with Transition Probabilities
Legend
Reliable Tracklet
S Source node
T Sink node
Cost-flow Network
S
T
Trajectory Completion
Tracking Output

Figure 2. Our proposed framework. In the cost-flow network, each node denotes a reliable tracklet; The flow costs of edges are defined by negative log of the affinity scores, which are obtained through a two-step target-specific metric learning and metric refinement processes on segments of short-time sequences known as local segments.

ing errors such as missed tracks or identity switches can be corrected. Then, similarity measurement between two track fragments (tracklets) to determine whether they belong to the same person becomes very critical in this tracklet association problem. Some of the state of art methods [20, 11] fuse several features such as motion, time, position, size and appearance to improve similarity measurement. However, their appearance models are still not adequate to handle large appearance variations, adversely affecting tracking performance.

As in [12], we advocate a discriminative target-specific appearance-based affinity model to reinforce the appearance cues for multi-person tracking. Unlike [12], we formulate the appearance model learning problem as a metric learning problem, which can provide reliable target-specific affinity scores between tracklets. Our target-specific metrics are online learned while the PIRMPT system proposed by [12] needs off-line learned local descriptors. Furthermore, our target-specific metrics are learned within each short-time segment known as a local segment instead of the entire sequence. This avoids the variability and complexity of learning metrics on a long-term sequence so that the additional computation due to learning can be greatly reduced. Due to less variability and complexity within a short-time segment, better metrics adaptive to the local segment can be learned more efficiently.

The framework proposed in this paper is shown in Figure 2. The target-specific metric learning process incorporates multi-person tracking cues, such as motion, spatio-temporal constraint and exit state. In contrast to conventional metric learning applications [14, 29, 15, 16], our proposed framework does not need off-line training. Moreover, we formulate the trajectory recovery by tracklet association problem as a min-cost flow network optimization. Each node in the network represents a reliable tracklet. The flow cost of a connected node pair is defined by a novel affinity score obtained through a two-step target-specific metric learning and metric refinement processes.

The main contribution of this paper is that a target-specific metric with strong discriminative power is online learned in two steps. The first step is to identify reliable tracklets. The second step estimates the affinity scores. The metrics are learned within each local segment for reduced computation and locally adaptive metrics.

The rest of this paper is organized as follows. Section 2 describes the cost-flow network formulation. Section 3 presents the online target-specific metric learning. Trajectory completion for full trajectory recovery is presented in section 4 . Experimental results and comparisons are shown in section 5. Section 6 concludes the paper.

2. Cost-flow Network Formulation for Trajectory Recovery by Tracklets Association

The cost-flow network has been shown to be effective for estimating the trajectories in the previous studies [28, 3, 19, 5]. However, in these works, the nodes are defined by the detection responses. In recent works [12, 25, 27, 26], tracklets were generated based on association of detection responses. We generate tracklets based on motion trajectory using successive shortest path algorithm. We can construct a smaller graph based on such tracklets which are of a higher order of abstraction than those based on detection responses. The problems in long-term tracking can be solved by directly linking tracklets instead of detection responses.

Given a video input, we first detect pedestrians in each frame by an existing detector, such as the DPM detector [7]. Similar to [20], we run the successive shortest path algorithm as in [19] which optimizes over trajectory to generate initial tracklets (track fragments). However, the initial tracklets may be unreliable because the detection responses in one tracklet may come from more than one person. Therefore, we use the online learned target-specific metrics to refine these initial tracklets for reliable tracklets. The cost-flow network formulation is based on the reliable tracklets and network flow optimization yields the trajectories of multiple persons.

We define an objective function for tracklet association which takes a similar form as detection association in [28]. Let $X = \{F_i\}$ be the collection of all the tracklets. A single trajectory hypothesis is defined as an ordered list of N tracklets: $T_k = \{F_{k_1}, F_{k_2}, ..., F_{k_l}\}$, where $F_{k_i} \in X$, and $i = 1, ..., l; 1 \leq l < N$. A tracklet association hypothesis $\mathcal{T}$ is defined as a set of single trajectory hypotheses: $\mathcal{T} = \{T_k\}$.

The objective of tracking association is to maximize the posteriori probability of $\mathcal{T}$ given X:

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{\mathcal{T}} P(\mathcal{T}|X) \\ &= \arg\max_{\mathcal{T}} P(X|\mathcal{T})P(\mathcal{T}) \\ &= \arg\max_{\mathcal{T}} \prod_i P(F_i|\mathcal{T})P(\mathcal{T}) \end{aligned} \tag{1}$$

assuming that the likelihood probabilities of F_i are conditionally independent.

If we assume that the motion of each tracklet is independent and one tracklet can only belong to one trajectory, we can further decompose the above equation into:

$$\mathcal{T}^* = \arg\max_{\mathcal{T}} \prod_i P(F_i|\mathcal{T}) \prod_{T_k \in \mathcal{T}} P(T_k) \tag{2}$$

$$s.t. \quad T_k \bigcap T_l = \Phi, \forall k \neq l \tag{3}$$

We define the second term in Equ. (2) as follows:

$$\begin{aligned} P(T_k) &= P(\{F_{k_1}, F_{k_2}, ..., F_{k_l}\}) \\ &= P_s(F_{k_1})(\prod_{n=1}^{l-1} P(F_{n+1}|F_n))P_t(F_{k_l}) \end{aligned} \tag{4}$$

$P(F_i|\mathcal{T})$ is the likelihood function of tracklet F_i. Here we assume that there are no false alarms from the reliable tracklets, so $P(F_i|\mathcal{T}) = 1$. Then Equ. (2) can be further simplified as follows:

$$\begin{aligned} \mathcal{T}^* &= \arg\max_{\mathcal{T}} \prod_i P(F_i|\mathcal{T}) \prod_{T_k \in \mathcal{T}} P(T_k) \\ &= \arg\max_{\mathcal{T}} \prod_{T_k \in \mathcal{T}} P(T_k) \end{aligned} \tag{5}$$

$P(T_k)$ is modeled as a Markov chain, which includes starting probability $P_s(F_{k_1})$, termination probability $P_t(F_{k_l})$, and transition probability $P(F_{n+1}|F_n)$ between temporarily adjacent tracklets. Finding the optimal association hypothesis $\mathcal{T}^*$ is equivalent to minimizing the cost of flow from source s to sink t in a network flow graph. A network graph can be constructed as follows:

Given an observation set X: for every tracklet $F_i \in X$, we create a node v_i, an edge from source s to a node,(s, v_i), with cost $c(s, v_i) = c_i^s$ and flow $f(s, v_i) = f_i^s$, and an edge from a node to sink t, (v_i, t) with cost $c(v_i, t) = c_i^t$ and flow $f(v_i, t) = f_i^t$. For every transition $P(F_j|F_i) \neq 0$, create an edge (v_i, v_j) , $i \neq j$, with cost $c(v_i, v_j) = c_{ij}$ and flow $f(v_i, v_j) = f_{ij}$. We take the logarithm of the objective function to simplify the expression while preserving the maximum a posteriori probability (MAP) solution. Then, Equ. (5) can be re-written as follows:

$$\mathrm{T} = \arg\min_{\mathcal{T}} \left(\sum_i c_i^s f_i^s + \sum_{ij} c_{ij} f_{ij} + \sum_i c_i^t f_i^t \right) \tag{6}$$

$$\begin{aligned} & s.t. \quad f_{ij}, f_i^s, f_i^t \in \{0, 1\}, \\ & and \quad f_i^s + \sum_j f_{ji} = f_i^t + \sum_j f_{ij} \end{aligned} \tag{7}$$

subject to Equ. (6), where

$$\begin{aligned} & c_i^s = -\log P_s(F_i), \quad c_i^t = -\log P_t(F_i), \\ & c_{ij} = -\log P(F_j|F_i). \end{aligned}$$

Equ. (7) ensures that the tracklet association hypothesis $\mathcal{T}$ is non-overlapping. The above formulation can be mapped into a cost-flow network with source s and sink t. Estimating the transition costs c_{ij} is the key factor in solving this min-cost network flow problem. Previous network flow approaches [28, 3, 19, 5] only utilize motion cues and simple appearance features such as RGB histograms to calculate c_{ij}. However, these cues are not very reliable when interactions and occlusions between targets with similar color appearance occur. Simple appearance models cannot handle large appearance variations. In this paper, we propose to learn target-specific segment-wise appearance-based model online for estimating c_{ij}.

3. Online Target-Specific Metric Learning for Tracklets Association

In this section, we introduce an online target-specific metric learning approach to obtain the affinity scores of adjacent tracklets, which can be used as the transition probabilities between two corresponding nodes in the cost-flow network. We perform local transition probabilities estimation within S frames ($S = 50$ in our implementation).

A novel target-specific appearance-based model is proposed to obtain effective appearance cues for reliable transition probabilities estimation. We formulate the appearance model learning problem as a metric learning problem, which can enhance the features with strong discriminative power and suppress the features with weak discriminative power. As a result, the learned models can better represent the appearance cues locally and provide reliable transition probabilities estimation. We learn target-specific metrics so that target-specific properties can be efficiently explored for more discriminative models. Contrasting to the work in [12] in which local descriptors are learned offline, our learning is online throughout and our target-specific metrics are adaptive to local segments.

3.1. Online Target-Specific Appearance-based Model Learning

We aim to online learn discriminative target-specific appearance-based models while keeping the computational complexity low. We learn the appearance model by formulating it as a metric learning problem. For each tracklet F_i, we learn a distance metric function.

The learning involves feature representation, online training sample collection and online training. To create a strong appearance-based model, we start from a rich set of basic features, which includes color, shape and texture, to describe a person's appearance. All the training images are normalized to 128×64 pixels. For the color feature, RGB, YCbCr and HSV color histograms are extracted with 16 bins for each channel respectively and concatenated into a 144-element vector. To capture shape information, we adopt the Histogram of Gradients (HOG) feature [6] by setting the cell size to be 8 to form a 3968-element vector. Two types of texture features are extracted by Schmid and Gabor filters. In total, 13 Schmid channel features and 8 Gabor channel features are obtained to form a 336-element vector by using a 16-bin histogram vector to represent each channel. Each person image is thus represented by a feature vector in a 4448-dimensional feature space.

Given a training dataset $Z = \{(z^t, l_i)\}_{i=1}^{n}$, where z^t is a 4448-dimensional feature vector representing the appearance of a detection response at frame t, and l_i is the tracklet label which the detection response belongs to. We define a positive difference vector x_i^p computed between a pair of relevant samples (detection responses belonging to the same person) and a negative difference vector x_i^n computed from a pair of irrelevant samples (detection responses belonging to different persons). Here, we assume the first M frames of each initial tracklets are reliable and the detection responses are from the same person. Training samples are therefore collected from these frames.

The difference vectors x_i^p and x_i^n are defined as follows:

$$\begin{aligned} x_i^p &= d(z_i, z_i') = |z_i - z_i'| \\ x_i^n &= d(z_i, z_j') = |z_i - z_j'|, \quad i \neq j \end{aligned} \tag{8}$$

where d is an absolute difference function, z_i and z_i' are two samples from the same tracklet F_i, z_j' is a sample from a different tracklet F_j.

Given the difference vectors x_i^p and x_i^n, a distance function D_i for tracklet F_i can be learned based on relative distance comparison so that $D_i(x_i^p) < D_i(x_i^n)$. This distance function D_i is parameterized as a Mahalanobis distance function:

$$D_i(x) = x^T M_i x, \quad M_i \succeq 0 \tag{9}$$

We adopt the logistic function as in [29] to learn D_i to force $D_i(x_i^p)$ to be small, and $D_i(x_i^n)$ to be big:

$$\min_{D_i} r(D_i) = -\log\left(\left(1 + \exp\left(D_i(x_i^p) - D_i(x_i^n)\right)\right)^{-1} \right) \tag{10}$$

The term M_i in the distance function D_i can be decomposed by eigendecomposition:

$$M_i = A_i \Lambda_i A_i^T = W_i W_i^T, \quad W_i = A_i \Lambda_i^{\frac{1}{2}} \tag{11}$$

where A_i is the orthonormal eigenvector matrix of M_i and the diagonal of Λ_i are the corresponding eigenvalues.

Hence, learning a distance function D_i is equivalent to learning the matrix W_i as follows:

$$\begin{aligned} &\min_{W_i} r(W_i), s.t. \quad w_i^T w_j = 0, \forall i \neq j, w_i, w_j \in W_i \\ &r(W_i) = \log(1 + \exp\{\|W_i^T x_i^p\|^2 - \|W_i^T x_i^n\|^2\}) \end{aligned} \tag{12}$$

Online training sample collection is an important issue in online learning. We use the q strongest ($q = 4$ in this work) detection responses in each tracklet as training samples. For x_i^p, we collect relevant sample pairs from the same tracklet. However, for x_i^n, we collect irrelevant sample pairs from different persons. To determine the relevance of sample pairs, two constraints are used: spatio-temporal and exit constraints. The first constraint is based on the fact that one person cannot appear at two or more different locations at the same time. The second constraint is based on the observation that the person who has already exited the view cannot be the person who is still within the view. We online collect irrelevant samples, which satisfy the above two constraints, from F_i and F_j respectively to form irrelevant pairs.

Once online training sample collection is finished, we adopt the gradient descent method to learn W_i for each tracklet F_i. Finally, we calculate the target-specific transform matrices for all the tracklets:

$$W = \{W_i\}, \quad i = 1, ..., N \tag{13}$$

3.2. Tracklet Refinement

To solve Equ. 6, we need to identify reliable tracklets for the nodes in the network. The initial tracklets are produced by the successive shortest path algorithm as in [19]. This method uses spatio-temporal information such as distance between corresponding observations in adjacent frames to link the detections into tracklets. Without effective use of appearance cues, the initial tracklets may be not consistent in appearance and hence unreliable when there is a lot of interactions or occlusions between targets. A typical error is that there are some detection responses belonging to different persons in one tracklet. Hence, tracklet refinement is needed to separate tracklets into multiple short but reliable ones.

We use the online learned target-specific metrics to refine the initial tracklets. To construct the probe set, the detection with the strongest detection response, g_i, is selected from the first M frames of an initial tracklet, F_i, which are assumed to be reliable. It is defined as $G = \{g_i\}$, $i = 1, ..., N_s$, where N_s are the number of tracklets in a local segment. And, each tracklet F_i has only one selected g_i in G.

We learn the target-specific transform matrix W_i for each initial tracklet after collecting training samples as described in previous sub-section. Then the identity test is carried out within a local segment frame by frame to obtain the relative distance between detection response at frame t of F_i and the corresponding g_i in the probe set:

$$\begin{aligned} x_i^t &= d(z_i^t - g_i) = |z_i^t - g_i|; \quad i = 1, ..., N_s \\ d_i^t &= \|W_i^T x_i^t\|^2 \end{aligned} \tag{14}$$

where z_i^t is one instance from tracklet F_i at frame t, g_i is the corresponding detection response of F_i in G, and d_i^t is the relative distance between z_i^t and g_i.

The relative distance between the current detection response z_i^t and the probe g_i for a reliable tracklet should be small; otherwise, it is an unreliable tracklet.

A distance threshold ω is used to determine reliable tracklets. In a tracklet F_i, if K ($K = 5$ in our implementation) consecutive detection responses having relative distance values (from g_i) above ω, we split F_i into two parts from the first consecutive detection response. We repeat the above process at multiple times until there are no unreliable tracklets.

3.3. Computing Tracklet Affinities

In this section, we present the calculation of the affinity score between F_i and F_j, or equivalently, the transition probability, P_{ij}, in the network between node i and node j. We first calculate the relative distances d_{ij}^t between each detection response in F_i and the probe g_j, and $d_{ji}^{t'}$ between each detection response in F_j and the probe g_i,

$$\begin{aligned} &x_{ij}^t = |z_i^t - g_j|, \; x_{ji}^{t'} = |z_j^{t'} - g_i|; \quad i, j = 1, ..., N_s \\ &d_{ij}^t = \|W_i^T x_{ij}^t\|^2, \; d_{ji}^{t'} = \|W_j^T x_{ji}^{t'}\|^2 \end{aligned} \tag{15}$$

where z_i^t denotes the feature vector of a detection response in tracklet F_i at frame t, $z_j^{t'}$ denotes the feature vector of a detection response in tracklet F_j at frame t', and $g_i, g_j \in G$, m, n are the number of frames of F_i and F_j respectively.

Then we calculate the mean values of the relative distances and use them to define affinity score, $\mathcal{S}_{ij}$, and hence equivalently, P_{ij}, as follows:

$$d_{ij} = (\sum_t d_{ij}^t)/m, \; d_{ji} = (\sum_{t'} d_{ji}^{t'})/n \tag{16}$$

$$\mathcal{S}_{ij} = (d_{ij} d_{ji})^{-1} \mathcal{C}_{ij} \tag{17}$$

where $\mathcal{C}_{ij}$ is a limiting function as explained below.

We do not have to apply Equ. (17) to every pair, because there are a lot of obviously non-related tracklet pairs which do not belong to the same person. To leave them out, a limiting function is proposed based on motion, spatio-temporal, and exit constraints:

$$\mathcal{C}_{ij} = C_m(F_i, F_j) C_t(F_i, F_j) C_e(F_i, F_j) \tag{18}$$

The motion constraint is defined by:

$$C_m(F_i, F_j) = \begin{cases} 1, & if \; |(p_i^{t_i^s} - p_j^{t_j^e})/\Delta t| < \sigma W_{z_i} \\ 0, & otherwise \end{cases} \tag{19}$$

where Δt is the time gap between the ending frame of F_j and the starting frame of F_i. W_{z_i} is the width of the instance z_i's detection window. σ is a threshold (σ=0.5 in our implementation).

This motion constraint C_m is based on the observation that if two tracklets belong to the same person, the velocity between the gap of the two tracklets is less than the width of either's detection window.

The spatio-temporal constraint is defined as follows:

$$C_t(F_i, F_j) = \begin{cases} 1, & if \; F_i \cap F_j = \phi \\ 0, & otherwise \end{cases} \tag{20}$$

where $\cap$ is an intersection operator that is used to find the overlap over time between two tacklets.

The exit constraint is defined as:

$$C_e(F_i, F_j) = \begin{cases} 1, & if\ t_i^s > t_j^e \ \&\ p_j^{t_j^e} \notin E \\ 0, & otherwise \end{cases} \tag{21}$$

where t_i^s is the starting frame of tracklet F_i, t_j^e is the ending frame of tracklet F_j, $p_j^{t_j^e}$ is the position of the detection response of tracklet F_j at time t_j^e and E is the exit area which is near image borders. For static cameras, we adopt the incremental learning algorithm for exit map as in [26] to obtain E.

$C_t(F_i, F_j)$ and $C_e(F_i, F_j)$ make the association between F_i and F_j becomes possible if they have no overlap in time and F_j does not exit the screen when F_i appears.

We obtain the transition costs of the adjacent nodes in the cost-flow network by taking negative logarithm of the affinity scores between corresponding tracklets:

$$c_{ij} = -\log \mathcal{S}_{ij} \tag{22}$$

Then we estimate the best tracklet association hypothesis $\mathcal{T}^*$ in Equ. (6) based on c_{ij}.

4. Trajectory Completion

After tracklet association, there are still some gaps between adjacent tracklets in each trajectory due to missed detections and occlusions. Hence, by imposing velocity continuity constraint over the gaps between tracklets, the trajectory over the gaps are estimated based on a linear motion model. The trajectory interpolation is also subject to the following two constraints: (1) the gaps is less than θ_f frames($\theta_f = 50$ in our implementation); and, (2) the corresponding affinity scores should be higher than a threshold θ_c.

5. Experiments

We present our experimental results in two sub-sections: comparison between our approach and several network flow based multi-tracking methods; and, comparison with other state-of-the-art methods on benchmark video sequences.

5.1. Comparison with other Network Flow Methods

The comparison is performed on the popular TUD Crossing sequence and ETH BAHNHOF sequence as in [5]. We use ID switches or the number of mismatches as the quantitative measures of performance, which is the same as in [5]. Table 1 shows ID switches and the total number of correct observations for the TUD Crossing sequence, and the first 350 frames of the ETH BAHNHOF sequence. We also show the tracking results of our approach without using the target-specific metric learning. We learn a common class metric for all the tracklets. This method is indicated as *baseline 1*. Note that our approach provides better results in all cases when compared with the three network flow methods [28, 19, 5]. Moreover, the obvious improvement in ID switches indicates that our approach can better deal with long-term tracking, where the traditional motion models are less reliable. Figure 3 shows the superiority of our approach.

Algorithm	TUD Crossing	ETH	ETH (GT)
DP [19]	32/768	37/1387	25/1648
LRMCNF [5]	14/819	23/1514	14/1783
MCNF [28]	9/433	11/1057	5/922
Baseline 1	10/845	5/1728	3/1786
Ours	7/862	1/1790	0/1820

Table 1. Comparison of tracking results with other network flow methods on TUD Crossing and ETH BAHNHOF (first 350 frames) sequences. The tracking results of the above three network flow based tracking methods are from [5]. The entries in the table are (ID switches)/(total number of correct observations used in the trajectories). Columns 1 and 2 use the pre-trained human detector of [7]. Column 3 shows the results when ground truth detections are used to generate the initial tracklets. The ground truth detections are from [19]. **DP algorithm of [19] is our baseline work without adding the online metric learning framework**. By comparison with [19], we can find that our approach improves a lot for both two metrics. *Baseline 1* is our approach using the common metric learning instead of the target-specific metric learning. From the results, we can see that our target-specific metric is more reliable than the common metric.

5.2. Comparison with state-of-the-art methods

We evaluate our approach on three public sequences: TUD Stadtmitte sequence, ETH BAHNHOF sequence and ETH SUNNY DAY sequence. To make fair comparisons, we use the same offline learned human detector [7] to detect person instances as in the compared methods. The quantitative comparisons are presented based on the widely used evaluation metrics [13] in this subsection. We use the evaluation codes downloaded from [17]. We also present the tracking results of our approach without using the target-specific metric learning as *baseline 1*. The superiority of our target-specific metric learning can be observed in these results.

For fair comparison, the experiments are conducted using the same TUD Stadtmitte dataset with groundtruth as defined in [27]. This video sequence, which contains 179 frames, is captured on a street at a relative low viewpoint and there are frequent occlusions and interactions among the pedestrians.

The tracking evaluation results are shown in Table 2.

Method	Recall	Precision	FAF	GT	MT	PT	ML	Frag	IDS
Energy Minimization [1]	-	-	-	9	60.0%	30.0%	0.0%	4	7
DC Tracking [2]	74.7%	84.2%	0.870	10	50.0%	50.0%	0.0%	8	10
PRIMPT [12]	81.0%	99.5%	0.028	10	60.0%	30.0%	10.0%	0	1
Online CRF Tracking [27]	87.0%	96.7%	0.184	10	70.0%	30.0%	0.0%	1	0
Baseline 1	95.1%	99.4%	0.030	10	100%	0.0%	0.0%	2	1
Ours	98.0%	99.3%	0.040	10	100%	0.0%	0.0%	3	0

Table 2. Comparison of tracking results between the state-of-art methods and ours on the TUD Stadtmitte dataset.

Note that our results are much better than those in [1, 2, 19]. Compared with [12, 27], the improvement is also obvious for some metrics. Our approach achieves the highest recall and the mostly tracked score (MT) among all the methods. It also achieves the lowest ID switches. Meanwhile, our approach achieves competitive performance on precision, false alarms per frame and fragments compared with [12, 27].

To show the effectiveness of our approach,we further evaluate our approach on the challenging ETH dataset [12], which is captured by a stereo pair of cameras mounted on a moving child stroller in a busy street scene. Because of the low view angle and forward moving cameras, occlusions and interactions of the targets frequently occur in these video sequences, which makes the dataset rather challenging.

We select the ”BAHNHOF” and ”SUNNY DAY” video sequences from ETH dataset used in [12, 19, 27] for experiments. The two sequences are both from the left camera and contain 999 and 354 frames respectively. For fair comparison with [12, 27, 18], we use the same groundtruth from [27] and no depth, structure-from-motion localization, and ground plane information is used.

The quantitative tracking results are shown in Table 3. We can see that our approach can achieve better or competitive performance on all the commonly used evaluation measures. Compared with [12], the most related work, the recall and precision are improved by 4.1% and 7.6% respectively; the MT is improved by 7.2%; false alarms per frame are reduced by 41.6%; and ID switches are reduced by 54.5%. The significant reduction in ID switches and false alarms indicates that our target-specific appearance-based model is superior to the method by [12].

5.3. Computation Speed

The computation speed depends on the number of targets in a video sequence. Our approach is implemented using MATLAB on a 3.3GHz PC with 8 GB memory. The average speed of our method is around 100 milli-seconds per frame for TUD and ETH datasets, respectively (excluding the detection and HOG feature extraction time). The codes can be further optimized.

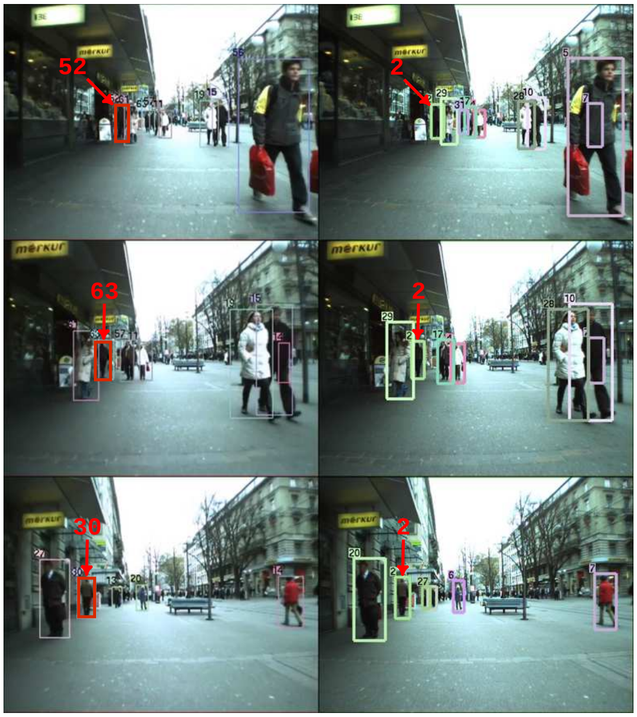


Figure 3. Columns 1 and 2 compare the tracking results of LRM-CNF [5] (left) and our approach (right) respectively on the ETH BAHNHOF sequence (frames 104 - first row, 130 -second row, 187- third row). Note the detection windows pointed by red arrows. We can see that our approach maintains ID labels more reliably.

6. Conclusion

We have proposed a novel introduction of online target-specific metric learning for trajectory recovery by tracklets association using network flow optimization for multi-person tracking. Instead of detection responses, tracklets are used as the nodes in the network graph, with edges defined by a cost computed from a novel tracklet affinity scores. The experimental results on four public sequences have shown significant improvements compared with state-of-art methods.

Method	Recall	Precision	FAF	GT	MT	PT	ML	Frag	IDS
PRIMPT [12]	76.8%	86.6%	0.891	125	58.4%	33.6%	8.0%	23	11
Online CRF Tracking [27]	79.0%	90.4%	0.637	125	68.0%	24.8%	7.2%	19	11
DTLE Tracking [18]	77.3%	87.2%	-	125	66.4%	25.4%	8.2%	69	57
Baseline 1	78.7%	92.0%	0.710	125	60.0%	29.6%	10.4%	77	19
Ours	80.9%	94.2%	0.520	125	65.6%	24.0%	10.4%	26	5

Table 3. Comparison of tracking results between the state-of-art methods and ours on the "BAHNHOF" and "SUNNY DAY" sequences.

References

[1] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization. In CVPR, 2011. 7

[2] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In CVPR, 2012. 1, 7

[3] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(9):1806–1819, 2011. 1, 2, 3

[4] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Online multi-person tracking-by-detection from a single, uncalibrated camera. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(9):1820–1833, 2011. 1

[5] A. Butt and R. T. Collins. Multi-target tracking by lagrangian relaxation to min-cost network flow. In CVPR, 2013. 1, 2, 3, 6, 7

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, 2005. 1, 4

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010. 1, 3, 6

[8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In CVPR, 2008. 1

[9] A. V. Goldberg. An efficient implementation of a scaling minimum-cost flow algorithm. Journal of Algorithms, 22:1–29, 1992. 1

[10] C. Huang and R. Nevatia. High performance object detection by collaborative learning of joint ranking of granule features. In CVPR, 2010. 1

[11] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In ECCV, 2008. 1, 2

[12] C. H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking? In CVPR, 2011. 2, 4, 7, 8

[13] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In CVPR, 2009. 1, 6

[14] J. Lu, J. Hu, X. Zhou, Y. Shang, Y. P. Tan, and G. Wang. Neighborhood repulsed metric learning for kinship verification. In CVPR, 2012. 2

[15] J. Lu, Y. P. Tan, and G. Wang. Discriminative multi-manifold analysis for face recognition from a single training sample per person. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1):39–51, 2013. 2

[16] J. Lu, G. Wang, and P. Moulin. Human identity and gender recognition from gait sequences with arbitrary walking directions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 9(1):51–61, 2014. 2

[17] A. Milan. Discrete-continuous optimization for multi-target tracking. http://www.gris.informatik.tu-darmstadt.de/~aandriye/dctracking.html. 6

[18] A. Milan, K. Schindler, and S. Roth. Detection- and trajectory-level exclusion in multiple object tracking. In CVPR, 2013. 7, 8

[19] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In CVPR, 2011. 1, 2, 3, 5, 6, 7

[20] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua. Tracking multiple people under global appearance constraints. In ICCV, 2011. 2, 3

[21] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In CVPR, 2007. 1

[22] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In CVPR, 2009. 1

[23] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In CVPR, 2005. 1

[24] J. Xing, H. Ai, and S. Lao. Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In CVPR, 2009. 1

[25] B. Yang, C. Huang, and R. Nevatia. Learning affinities and dependencies for multi-target tracking using a crf model. In CVPR, 2011. 2

[26] B. Yang and R. Nevatia. Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In CVPR, 2012. 2, 6

[27] B. Yang and R. Nevatia. An online learned crf model for multi-target tracking. In CVPR, 2012. 2, 6, 7, 8

[28] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In CVPR, 2008. 1, 2, 3, 6

[29] W. S. Zheng, S. Gong, and T. Xiang. Re-identification by relative distance comparison. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(3):653–668, 2013. 2, 4