How does Person Identity Recognition Help Multi-Person Tracking?

Cheng-Hao Kuo and Ram Nevatia
University of Southern California, Institute for Robotics and Intelligent Systems
Los Angeles, CA 90089, USA
{chenghak|nevatia}@usc.edu

Abstract

We address the problem of multi-person tracking in a complex scene from a single camera. Although tracklet-association methods have shown impressive results in several challenging datasets, discriminability of the appearance model remains a limitation. Inspired by the work of person identity recognition, we obtain discriminative appearance-based affinity models by a novel framework to incorporate the merits of person identity recognition, which help multi-person tracking performance. During off-line learning, a small set of local image descriptors is selected to be used in on-line learned appearances-based affinity models effectively and efficiently. Given short but reliable tracklets generated by frame-to-frame association of detection responses, we identify them as query tracklets and gallery tracklets. For each gallery tracklet, a target-specific appearance model is learned from the on-line training samples collected by spatio-temporal constraints. Both gallery tracklets and query tracklets are fed into hierarchical association framework to obtain final tracking results. We evaluate our proposed system on several public datasets and show significant improvements in terms of tracking evaluation metrics.

1. Introduction

Tracking multiple people in a real scene is an important topic in the field of computer vision since it has many applications such as surveillance systems, robotics, and human-computer interaction environments. This is a highly challenging problem, especially in complex and crowded environments with frequent occlusions and interactions of targets. For example, Figure 1 shows several busy scenes which are difficult cases for multi-person tracking.

In recent literature, human detection techniques have achieved impressive progress [7, 10, 13, 24, 25, 26], which enable a popular tracking scheme: tracking by tracklet-association [14, 17, 28, 29]. The main idea is to link detection responses or short tracklets gradually into longer tracks by optimizing linking scores or probabilities between tracklets globally. Considering the information from future frames, some detection errors such as missed detections and false alarms can be corrected; this also solves the problem of frequent or long occlusions between targets effectively. A key element in tracklet-association is the affinity measurement between two tracklets to decide whether they belong to the same target or not. Relying on spatio-temporal restrictions in video sequences, these types of methods usually fuse several features such as motion, time, position, size, and appearance as the affinity measurement. However, due to the computational constraints, often in previous work only simple features are used as their appearance models, which limit the accuracy.

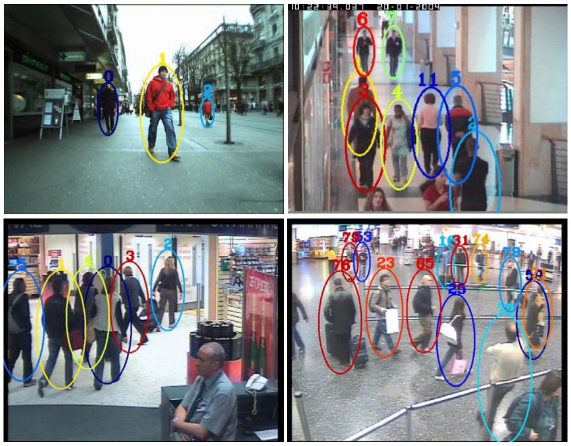


Figure 1. Some snapshots of videos from our multi-person tracking results. The goal of this work is to locate the targets and maintain their identities in a real and complex scene.

To enhance the human appearance model, we explore another interesting topic: appearance-based person identity recognition [9, 11, 12, 20, 22]. Given an image of a person, the recognition system finds the best match in the gallery set. In fact, multi-person tracking and person recognition can be viewed as highly connected tasks since solving the problem of whether two tracklets are from the same person is essentially the recognition problem. Nevertheless, there

	Multi-person Tracking	Person Recognition
Appearance models	simple	complex
View-point change	small	large
Illumination change	small	large
Gallery size	small	large
Other cues	yes	no
Occlusion handling	yes	no

Table 1. Comparison between multi-person tracking and person identity recognition.

exist some differences between the approaches of these two problems. Compared to multi-person tracking where several cues can be applied, person recognition methods typically use the appearance as the only cue to help the association between a query and a gallery of people. Besides, to deal with the larger view-point and illumination changes, the appearance model used for person recognition is, in general, more complex than that in multi-person tracking. Some comparisons between these two problems are listed in Table 1.

In this paper, we propose a framework to incorporate the merits of person identity recognition to help multi-person tracking performance. Compared to normal application of person recognition, for multi-person tracking we have a small, though dynamic gallery. The proposed system is named as Person Identity Recognition based Multi-Person Tracking (PIRMPT). A recent work [15] used tracking by tracklet-association method and proposed a strategy to collect training samples for learning on-line discriminative appearance models (OLDAMs). PIRMPT adopts a similar framework but makes significant improvements with several new ideas. Unlike [15] which uses pre-defined local image descriptors, in PIRMPT a set of most discriminative features is selected by automatic learning from a large number of local image descriptors. This set serves as the feature pool for on-line learning of appearance-based affinity models. In the test phase, tracklets formed by frame-to-frame association are classified as query tracklets or gallery tracklets. For each gallery tracklet, a target-specific appearance-based affinity model is learned from the on-line training samples collected by spatio-temporal constraints, instead of learning a global OLDAMs for all tracklets in [15]. Both gallery tracklets and query tracklets are then fed into a hierarchical association framework to obtain final tracking results. The block diagram of the PIRMPT system is shown in Figure 2. We evaluate our proposed method on several public datasets: CAVIAR, TRECVID 2008, and ETH to show significant improvements in terms of tracking evaluation metrics.

The rest of the paper is organized as follows. Related work is discussed in Section 2. The definition of appearance-based affinity models is presented in Section 3. Tracklet association framework is presented in Section 4. The experimental results are shown in Section 5. The conclusion is given in Section 6.

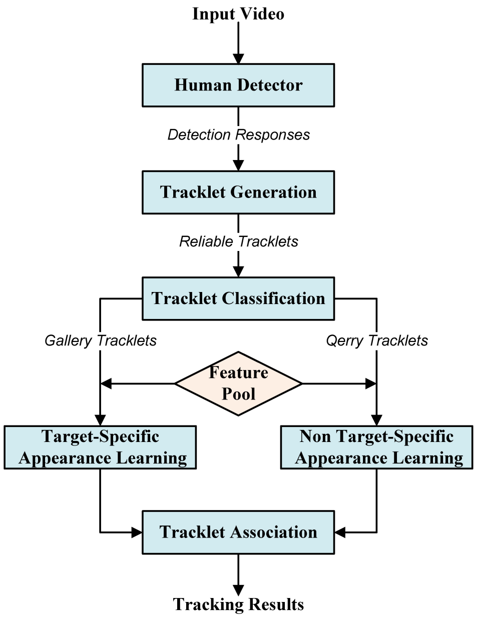


Figure 2. The block diagram of our proposed method.

2. Related Work

Due to the impressive advances in object detection [7, 10, 13, 24, 25, 26], detection-based tracking methods have gained increasing attention since they are essentially more robust in complex environments, even when the camera is moving. One typical solution is to adopt a particle filtering framework for representing the tracking uncertainty in a Markovian manner by only considering detection responses from past frames. Okuma *et al.* [19] formed the proposal distribution of the particle filter from a mixture of the Adaboost detections and the dynamic model. Cai *et al.* [5] worked on the same dataset as in [19] and made the improvement by rectification technique and mean-shift embedded particle filter. Breitenstein *et al.* [4] used the continuous confidence of pedestrian detectors and online trained classifiers as a graded observation model. These methods are suitable for online applications since the results are based on the past and current frames; no information from future frames are considered. However, it may fail when long occlusion of targets exists and sensitive to imperfect detection responses.

In contrast to those methods which only consider the past information, several approaches are proposed to optimize multiple trajectories globally, *i.e.* by considering the future frames in a time sliding window. Leibe *et al.* [16]

used Quadratic Boolean Programming to couple the detection and estimation of trajectory hypotheses. Andriluka *et al.* [3] applied Viterbi algorithm to obtain optimal object sequences. Zhang *et al.* [29] used a cost-flow network to model the MAP data association problem. Xing *et al.* [28] combined local tracklets filtering and global tracklets association. Huang *et al.* [14] proposed a hierarchical association framework to link shorter tracklets into longer tracks. Li *et al.* [17] adopted similar structure as [14] and presented a HybridBoost algorithm to learn the affinity models between two tracklets. The underlying philosophy is that observing more frames before making association decisions should generally help overcome ambiguities caused by long term occlusions and false or missed detections.

On the other hand, person identity recognition is a less addressed problem. Unlike tracking where motion and position are used to help maintain identities of targets, appearance is the only cue in person identity recognition. Gheissari *et al.* [11] developed two person re-identification approaches which use interest operators and model fitting for establishing spatial correspondences between individuals. Gray *et al.* [12] presented a method of performing viewpoint invariant pedestrian recognition using the ensemble of localized features. Farenzena *et al.* [9] found the asymmetry/symmetry axes and extracted the symmetry-driven accumulation of local features. Schwartz *et al.* [22] established a high-dimensional signature which is then projected into a low-dimensional discriminant latent space by Partial Least Squares reduction. Oreifej *et al.* [20] proposed a system to recognize humans across aerial images by Weighted Region Matching (WRM).

3. Appearance-based affinity models

Appearance models play an important role in both person recognition and multi-person tracking. Previous methods for person recognition usually propose complex features or a large number of image descriptors which need extensive computational power. On the other hand, multi-person tracking often use simple features such as color histograms for the speed issue. We aim to construct the appearance models which have strong discriminative power, yet are efficient to fulfill the speed requirements for the problem of multi-person tracking.

3.1. Local image descriptors and similarity measurements

To establish a strong appearance model, we extract a rich set of local descriptors to describe a person's image. A local descriptor d consists of a feature channel λ and a support region r, where $\lambda = (Color, Shape, Texture)$ and $r = (x, y, w, h)$. Given an image sample I, a single descriptor $d_{i,j}$ extracted over r_j via λ_i is denoted as

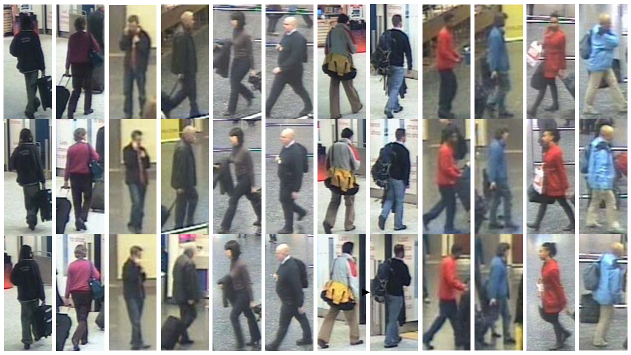


Figure 3. The off-line training samples for appearance models. Images in each column indicate the same person.

$$d_{i,j} = I(\lambda_i, r_j) \tag{1}$$

where i and j are the indices of the feature channel and the support region respectively.

In our design, the support regions $\{r\}$ are sampled from a large set of all possible rectangles within I, with the constraints of the width to height ratio fixed to 1:1, 1:2, or 2:1, which gives us 654 rectangles. For the feature channel, RGB color histograms are used with 8 bins for each channel and concatenated into a 24-element vector. To capture shape information, we adopt the Histogram of Gradients (HOG) feature [7] by concatenating 8 orientations bins in 2×2 cells over r to form a 32-element vector. To describe the image texture, we use a descriptor based on covariance matrices of image features proposed in [23].

Given the appearance descriptors, we can compute similarity between two person image patches. The color histogram and HOG feature are histogram-based features so standard measurements, such as χ^2 distance, Bhattacharyya distance, and correlation coefficient can be used; we choose correlation coefficient for simplicity. Distance between two covariance matrices is determined by solving a generalized eigenvalue problem, as described in [23]. The similarity score s between two image patches based on a certain local descriptor can be written as:

$$s_{i,j} = \rho_i\big(I_1(\lambda_i, r_j), I_2(\lambda_i, r_j)\big) \tag{2}$$

where ρ_i is the corresponding similarity measurement function of feature channel λ_i.

3.2. Model definition and descriptor selection

We define the appearance-based affinity model as an ensemble of local descriptors and their corresponding similarity measurements. It takes any two images of persons as input and computes an affinity score as the output. The desired model has the goal of discriminating between the

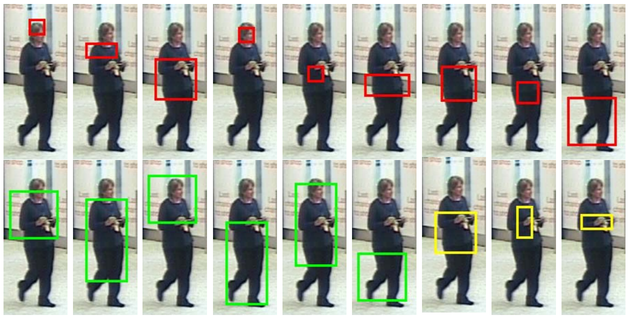


Figure 4. Some sample feature selected by Adaboost algorithm. The local descriptors of color histograms, HOG, and covariance matrices are indicated by red, green, and yellow respectively.

correct and the wrong pairs. The larger the affinity score is, the more likely it is that the two images belong to the same person. We design the appearance-based affinity models to be a linear combination of all similarity measurements by Equation 2. It takes the following form:

$$H(I_1, I_2) = \sum \alpha_{i,j} s_{i,j} \tag{3}$$

where the coefficients $\{\alpha\}$ represent the importance of local descriptors.

The training data are obtained from the tracking ground truth of TECVID 2008 dataset [2] provided by [17]. For each individual, M images of that person are extracted randomly along its trajectory. Some examples are provided in Figure 3. A training sample for the learning algorithm is defined as a pair of images: a positive sample is collected by a pair of images from the same person; a negative one is collected by a pair of images from any two different persons. The similarity scores of training samples are used as features in a learning framework.

We may use a large set of local descriptors via different channels and support regions. However, if we plan to include all of them in on-line multi-person tracking, the computation cost would be too high. Hence, we apply the standard Adaboost algorithm [21] to sequentially learn the features which are effective in comparing two images; it may be regarded as a feature selection process. This selected smaller set becomes the feature pool for the use of multi-person tracking, as the diamond block in Figure 2. In general, the color histograms are selected most; the covariance matrices are selected least. Color histograms tend to have smaller regions; HOG features tend to have larger regions. Some sample features are shown in Figure 4.

4. Tracklet association framework

As in Figure 2, PIRMPT system involves four parts: tracklet generation, tracklet classification, on-line appearance-based affinity models learning, and tracklet association. We describe each important component in this section.

4.1. Reliable Tracklets

In a given time sliding window, we apply a state-of-the-art human detector such as [7, 10, 13] on each frame. A dual-threshold association strategy is applied to detection responses in consecutive frames and generate short but reliable tracklets as in [14]. Based on the assumption that targets have small displacements over neighboring frames in a video sequence, we form a affinity score matrix S where each element in S is defined as the product of three similarity scores based on position, size and appearance between r_m and r_n as in [27]:

$$S(m, n) = A_{pos}(r_m, r_n) A_{size}(r_m, r_n) A_{appr}(r_m, r_n) \tag{4}$$

Two responses r_m and r_n in two neighboring frames are linked if their affinity score $S(m, n)$ is higher than a threshold θ_1 and exceeds any other elements in the m-th row and n-th column of S by another threshold θ_2. This strategy is conservative and biased to link only reliable associations between any two consecutive frames.

Since the detection responses in a generated tracklet may be noisy and not well-aligned to the target, tracklet refinement is needed to extract correct descriptors of motion and appearance. Let x_k indicate the the position and the size of an observation r_k in a tracklet $T = \{r_k\}$, where k is the index of time frame. We define the probability of a certain state $\{\hat{x}_k\}$ given the observation $\{x_k\}$ as:

$$P(\{\hat{x}_k\}|\{x_k\}) = \prod_k G(x_k - \hat{x}_k, \Sigma_p) \prod_k G(v_k, \Sigma_v) \prod_k G(a_k, \Sigma_a) \tag{5}$$

where $v_k = \frac{\hat{x}_{k+1} - \hat{x}_k}{t_{k+1} - t_k}$ and $a_k = \frac{v_k - v_{k-1}}{0.5(t_{k+1} - t_{k-1})}$ are the velocity and the acceleration at frame t_k; $G(., \Sigma)$ is the zero-mean Gaussian distribution. For each tracklet, the estimation of the true states $\hat{x}_k$ can be computed as:

$$\{\hat{x}_k\}^* = \arg\max_{\{\hat{x}_k\}} P(\{\hat{x}_k\}|\{x_k\}) \tag{6}$$

4.2. Tracklets classification

How to learn effective appearance-based affinity models is a key problem for robust performance in multi-person tracking. [15] proposed an approach to learn the global appearance-based affinity model which is shared by all tracklets in a given time sliding window, *i.e.*, all tracklets use the same model. In contrast to this work, we plan to learn target-specific models.

Inspired by the work of person identity recognition, we try to include the concept of gallery and query, with some necessary modifications which can be incorporated in multi-person tracking. We divide all tracklets into two groups: ”gallery” tracklets and ”query” tracklets. Different groups of tracklets have different strategies to learn their own appearance-based affinity models. In our design, a tracklet which is considered as a gallery tracklet needs to fulfill two requirements: 1) it is longer than a certain threshold; 2) it is not totally or heavily occluded by any other tracklet. The first requirement is based on the observation that a longer tracklet is more likely to be a true tracklet of a person; a shorter tracklet tends to be false alarm so that it is not appropriate to be registered as a gallery tracklet. The reason for second requirement is that the heavily occluded tracklets are not suitable to extract the local descriptors. Given the optimal states of each tracklet in Equation 6, an occupancy map is established in every frame and the visibility ratio is computed for each x_k of each tracklet. The second requirement is satisfied if the number of frames with the visibility ratio less than a certain threshold θ_v, is larger than $M(\theta_v = 0.7$, $M = 8$ in our implementation). If a tracklet is not classified as a gallery tracklet, it will be considered as a query tracklet.

4.3. On-line appearance-based affinity models

The construction of on-line appearance-based affinity models contains three parts: local descriptors extraction, on-line training sample collection, and the learning framework. A small set of discriminative local descriptors is learned off-line as described in Section 3. For each tracklet, we extract these local descriptors from the refined detection responses in head part (the first M frames) and tail part (the last M frames). Based on the occupancy map, we do not extract local descriptors for the detection responses whose visibility ratio is less than θ_v.

On-line training sample collection is an important issue in the learning of target-specific appearance model. We adopt assumptions similar to those in [15]: 1) detection responses in one tracklet are from the same person; 2) any detection responses in two different tracklets which overlap in time are from different persons. The first assumption is based on the observation that dual-threshold strategy generates reliable tracklets; the second one is based on the fact that one person can not appear at two different locations at the same time. For a certain tracklet T_i, a conflict set $\mathcal{C}_i$ is established where no element in $\mathcal{C}_i$ can be the same person as T_i. Next, we extract any two different detection responses from T_i as positive training samples, and two responses from T_i and $T_j \in \mathcal{C}_i$ respectively as negative training samples. The training set for tracklet T_i is denoted by $\mathcal{B}_i$.

For a gallery tracklet $T_i^{gallery}$, the training data $\mathcal{B}_i$ is used to learn a target-specific appearance-based affinity model. For a query tracklet T_i^{query}, the training data contain the union of all $\mathcal{B}_i$ in our design since a query tracklet is less robust to learn a meaningful target-specific model.

Once on-line training sample collection is finished, we compute the similarity scores between local appearance descriptors via all feature channels over all support regions. These similarity measurements serve as weak learners which are used in a standard boosting algorithm, as in [21], to learn the weight coefficients $\{\alpha\}$.

4.4. Tracklets association

We adopt the hierarchical tracklet association framework of [14]. The linking probability between two tracklets is defined by three cues: motion, time, and appearance:

$$P_{link}(T_i, T_j) = A_m(T_i, T_j) A_t(T_i, T_j) A_{appr}(T_i, T_j) \quad (7)$$

The motion model is defined by

$$\begin{aligned} A_m(T_i, T_j) = & G(x_i^{tail} + v_i^{tail}\Delta t - x_j^{head}, \Sigma_j) \cdot \\ & G(x_j^{head} - v_j^{head}\Delta t - x_i^{tail}, \Sigma_i) \end{aligned} \quad (8)$$

where Δt is the time gap between the tail of T_i and the head of T_j; x_i and v_i are refined positions and velocities of the head part or tail part of T_i.

The time model is simply a step function

$$A_t(T_i, T_j) = \begin{cases} 1, & \text{if } \Delta t > 0 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

which makes the link between T_i and T_j becomes possible if the tail of T_i appears earlier than the head of T_j.

The appearance-based affinity model is a linear combination of several similarity measurements of a set of local descriptors, as mentioned in Section 3. Every gallery tracklet has its own appearance-based affinity model: *i.e.* target-specific weighted coefficients $\{\alpha\}_i^{gallery}$. All query tracklets shared the same global weighted coefficients $\{\alpha\}^{query}$. The appearance-based affinity models used in $A_{appr}(T_i, T_j)$ depends on whether T_i or T_j is a gallery tracklet or not. $\{\alpha\}_i^{gallery}$ is used if T_i is a gallery tracklet; $\{\alpha\}_j^{gallery}$ is used if T_j is a gallery tracklet but T_i is not. $\{\alpha\}^{query}$ is used if T_i and T_j are both query tracklets.

5. Experimental results

To evaluate the performance of the PIRMPT, experiments are conducted on three public datasets: the CAVIAR Test Case Scenarios [1], the TRECVID 2008 [2], and the ETH Mobile Platform [8] dataset. The comparison between PIRMPT and several state-of-the-art methods is given based on the commonly used evaluation metrics [17]. Computation speed is also provided.

5.1. CAVIAR dataset

The CAVIAR Test Case Scenarios dataset was captured in a shopping center corridor by two fixed cameras from two different viewpoints. We use the video clips from the

Method	Recall	Precision	FAF	GT	MT	PT	ML	Frag	IDS
Wu *et al.* [27]	75.2%	-	0.281	140	75.7%	17.9%	6.4%	35*	17*
Zhang *et al.* [29]	76.4%	-	0.105	140	85.7%	10.7%	3.6%	20*	15*
Xing *et al.* [28]	81.8%	-	0.136	140	84.3%	12.1%	3.6%	24*	14*
Huang *et al.* [14]	86.3%	-	0.186	143	78.3%	14.7%	7.0%	54	12
Li *et al.* [17]	89.0%	-	0.157	143	84.6%	14.0%	1.4%	17	11
Kuo *et al.* [15]	89.4%	96.9%	0.085	143	84.6%	14.7%	0.7%	18	11
PIRMPT	88.1%	96.6%	0.082	143	86.0%	13.3%	0.7%	17	4

Table 2. Comparison of tracking results between the state-of-the-arts and PIRMPT on CAVIAR dataset. *The numbers of Frag and IDS in [27] [29] [28] are obtained by looser evaluation metrics. The human detection results we use are the same as [14, 17, 15].

Method	Recall	Precision	FAF	GT	MT	PT	ML	Frag	IDS
Huang *et al.* [14]	71.6%	80.8%	-	919	57.0%	28.1%	14.9%	487	278
Li *et al.* [17]	80.0%	83.5%	-	919	77.5%	17.6%	4.9%	310	288
Kuo *et al.* [15]	80.4%	86.1%	0.992	919	76.1%	19.3%	4.6%	322	224
PIRMPT	79.2%	86.8%	0.920	919	77.0%	17.7%	5.2%	283	171

Table 3. Comparison of tracking results between the state-of-the-arts and PIRMPT on the TRECVID 2008 dataset. The human detection results we use are the same as [14, 17, 15].

corridor view only. The frame rate is 25 fps and the image size is 384×288 pixels for all videos. For fair comparison with other state-of-the-art methods, the experiments are conducted on 20 clips[1] selected by [29]. There are in total 143 people across 29,283 frames and 77,270 detection annotations in the ground truth of 20 videos. The tracking evaluation results are shown in Table 2. Note that the ID switches are reduced by 64% by the PIRMPT approach; it also achieves lower fragmentation and false alarms per frame compared to other methods while keeping competitive numbers on recall rate, precision rate, and mostly tracked trajectories. Some sample frames with tracking results are presented in Figure 5(a).

5.2. TRECVID 2008 dataset

The TRECVID 2008 event detection dataset contains videos from with 5 fixed cameras covering different field-of-views in an airport. Authors in [17] extracted 9 videos from three cameras (Cam 1, Cam 3, and Cam 5) and annotated the tracking grounds truth for their evaluation on multi-target tracking. Each video features the 25 fps of frame rate, 720×576 pixels of the image size, and 5000 frames in length. In the ground truth, there are in total 919 people across 45,000 frames and 342,814 detection annotations from 9 videos. TRECVID 2008 dataset is much more difficult compared to the CAVIAR dataset, due to its high crowd density, heavy occlusions, and frequent interactions between different targets. Table 3 presents a comparison with results from [14, 17, 15], which have shown the impressive results on this challenging dataset. Our proposed method decreases the fragments and ID switches significantly. Compared to [15], the fragmentation is reduced 12% and the ID switch is reduced 24%. It shows that the person identity recognition helps the tracking system make correct associations between tracklets and improve the tracking results. Some sample frames with tracking results are presented in Figure 5(b,c).

5.3. ETH mobile platform dataset

The ETH dataset [8] was captured by a stereo pair of forward-looking cameras mounted on a moving children's stroller in a busy street scene. Due to the lower position of the cameras, total occlusions happen often in these videos, which increases the difficulty of this dataset. The frame rate is 13∼14 fps and the image size is 640×480 pixels for all videos. The ground truth annotations provided in the website[2] of [8] is only for pedestrian detection, not for multi-person tracking. Several previous papers [8, 6, 18] reported their tracking results based on mostly tracked trajectories, fragments, and ID switches by manual counting, which is time-consuming and prone-to-error. To avoid this, we created our tracking ground truth for automatic evaluation; two sequences "BAHNHOF" and "SUNNY DAY" from left camera only are used for our experiments. In our annotation, "BAHNHOF" sequence contains 95 individuals over 999 frames; "SUNNY DAY" sequence contains 30 in-

[1] Originally there are 26 videos in CAVIAR dataset. The selected 20 videos are EnterExitCrossingPaths1, EnterExitCrossingPaths2, OneLeaveShop1, OneLeaveShopReenter2, OneShopOneWait1, OneShopOneWait2, OneStopEnter1, OneStopEnter2, OneStopMoveEnter1, OneStopMoveEnter2, OneStopMoveNoEnter1, OneStopNoEnter1, OneStopNoEnter2, ShopAssistant1, ShopAssistant2, ThreePastShop1, TwoEnterShop1, TwoEnterShop3, TwoLeaveShop2, and WalkByShop1.

[2] http://www.vision.ee.ethz.ch/∼aess/dataset/

Sequence	Recall	Precision	FAF	GT	MT	PT	ML	Frag	IDS
BAHNHOF	76.5%	86.6%	0.976	95	51	37	7	21	10
SUNNY DAY	77.9%	86.7%	0.653	30	22	5	3	2	1

Table 4. Tracking results from PIRMPT on sequences ”BAHNHOF” and ”SUNNY DAY” from the ETH dataset.

dividuals over 354 frames. No stereo depth maps, structure-from-motion localization, and ground plane estimation are utilized in our method.

The tracking result are presented in Table 4. Note that Mitzel *et al*. [18] proposed a segmentation-based tracker and reported their numbers on the same two sequences recently. However, direct comparison is difficult as the evaluation in [18] is based on manual counting. The number of targets in our ground truth is much larger than theirs because we include smaller targets as well as short tracks to have a complete annotation. Besides, people who undergo long and total occlusion and then appear again are still considered the same persons in our case. In this type of situation, the appearance models enhanced by person identity recognition is more important when the motion information is not reliable. Some sample frames with tracking results are presented in Figure 5(d,e).

5.4. Speed

The computation speed of tracking depends on the density of observations from the detector, *i.e*. the number of detection responses which are present in each frame in the videos. The execution time are measured on 20 videos from the CAVIAR dataset and 9 videos from the TRECVID 2008 dataset. On average, the runtime speed is 48 fps for CAVIAR and 7 fps for TRECVID 2008, not including the processing time of the human detector. Our implementation is coded in C++ on a Intel Xeon 3.0 GHz PC.

6. Conclusion

We present a system, PIRMPT, to combine the merits of person recognition to help the performance of multi-person tracking. By off-line learning, a small number of local image descriptors is selected to be used in tracking framework for maintaining the effectiveness and efficiency. Given reliable tracklets, we identify them as query tracklets and gallery tracklets. For each gallery tracklet, a target-specific appearance-based affinity model is learned from the on-line training samples collected by spatio-temporal constraints. Experiments on challenging datasets show significant improvements by our proposed PIRMPT.

Acknowledgments

This paper is based upon work supported in part by Office of Naval Research under grant number N00014-10-1-0517.

References

[1] Caviar dataset. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.

[2] National institute of standards and technology: Trecvid 2008 evaluation for surveillance event detection. http://www.nist.gov/speech/tests/trecvid/2008/.

[3] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *CVPR*, 2008.

[4] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. Robust tracking-by-detection using a detector confidence particle filter. In *ICCV*, 2009.

[5] Y. Cai, N. de Freitas, and J. J. Little. Robust visual tracking for multiple targets. In *ECCV*, 2006.

[6] W. Choi and S. Savarese. Multiple target tracking in world coordinate with single, minimally calibrated camera. In *ECCV*, 2010.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[8] A. Ess, B. Leibe, K. Schindler, , and L. van Gool. A mobile vision system for robust multi-person tracking. In *CVPR*, 2008.

[9] M. Farenzena, L. Bazzani, A. Perina1, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010.

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

[11] N. Gheissari, T. B. Sebastian, P. H. Tu, and J. Rittscher. Person reidentification using spatiotemporal appearance. In *CVPR*, 2006.

[12] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *ECCV*, 2008.

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

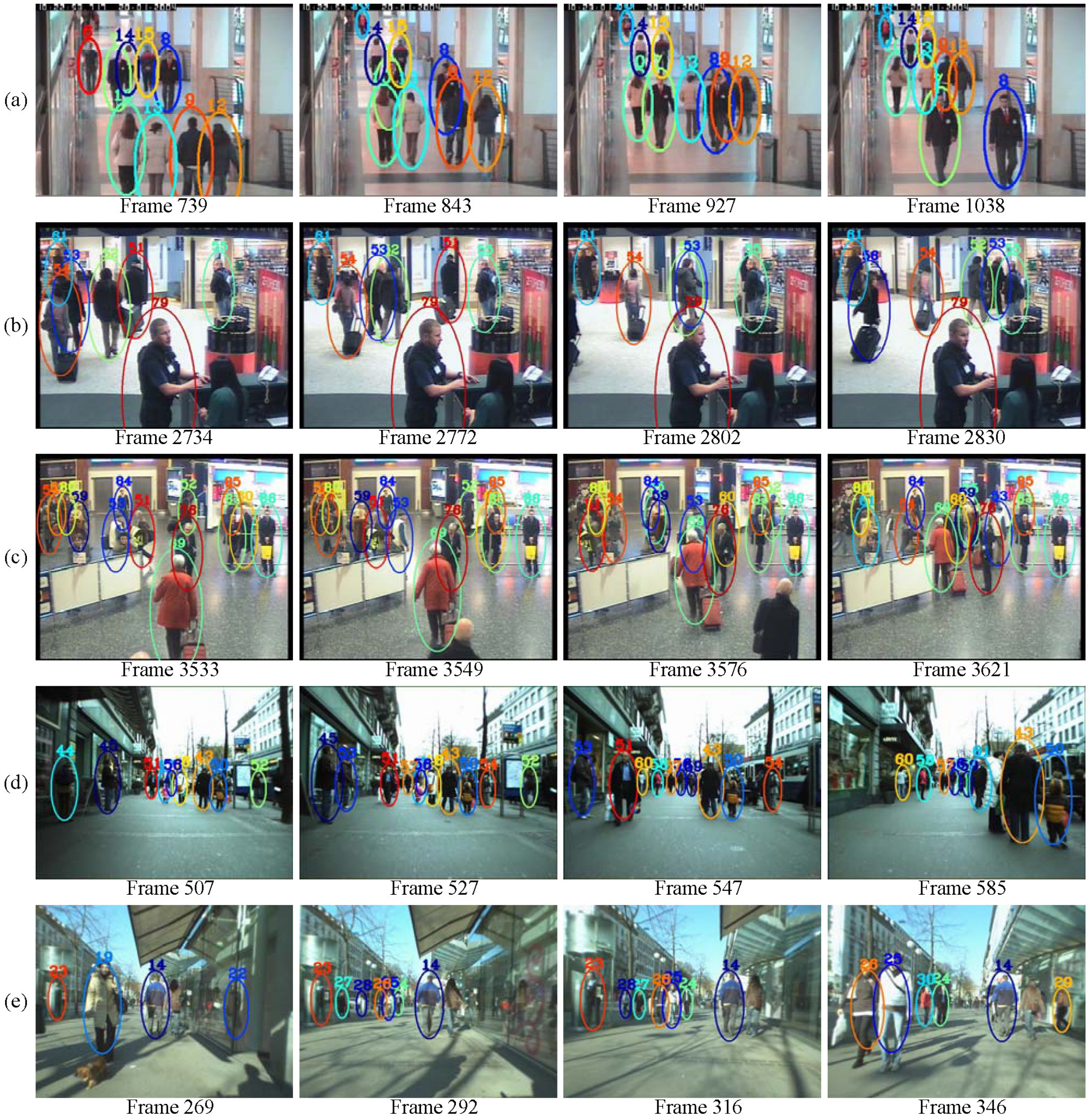
[14] C. Huang, B. Wu, and R. Nevatia. Robust object tracking by hierarchical association of detection responses. In *ECCV*, 2008.

[15] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, 2010.

[16] B. Leibe, K. Schindler, and L. V. Gool. Coupled detection and trajectory estimation for multi-object tracking. In *ICCV*, 2007.

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

[18] D. Mitzel, E. Horbert, A. Ess, and B. Leibe. Multi-person tracking with sparse detection and continuous segmentation. In *ECCV*, 2010.



Frame 739 Frame 843 Frame 927 Frame 1038

Frame 2734 Frame 2772 Frame 2802 Frame 2830

Frame 3533 Frame 3549 Frame 3576 Frame 3621

Frame 507 Frame 527 Frame 547 Frame 585

Frame 269 Frame 292 Frame 316 Frame 346

Figure 5. Sample tracking result on (a) CAVIAR, (b,c) TRECVID 2008, and (d,e) ETH dataset.

[19] K. Okuma, A. Taleghani, O. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *ECCV*, 2004.

[20] O. Oreifej, R. Mehran, and M. Shah. Human identity recognition in aerial images. In *CVPR*, 2010.

[21] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. In *Machine Learning*, pages 80–91, 1999.

[22] W. R. Schwartz and L. S. Davis. Learning discriminative appearance-based models using partial least squares. In *SIBGRAPI*, 2009.

[23] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *ECCV*, 2006.

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

[25] X. Wang, T. X. Han, and S. Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, 2009.

[26] 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.

[27] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors. *International Journal of Computer Vision(IJCV)*, 75(2):247–266, November 2007.

[28] 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.

[29] L. Zhang, Y. Li, and R. Nevatia. Global data association for multi-object tracking using network flows. In *CVPR*, 2008.