# Fusing Appearance and Spatio-temporal Features for Multiple Camera Tracking

Nam Trung Pham, Karianto Leman, Richard Chang,
Jie Zhang, and Hee Lin Wang

Institute for Infocomm Research, Singapore
{ntpham,karianto,rpchang,zhangj,hlwang}@i2r.a-star.edu.sg

**Abstract.** Multiple camera tracking is a challenging task for many surveillance systems. The objective of multiple camera tracking is to maintain trajectories of objects in the camera network. Due to ambiguities in appearance of objects, it is challenging to re-identify objects when they re-appear in other cameras. Most research works associate objects by using appearance features. In this work, we fuse appearance and spatio-temporal features for person re-identification. Our framework consists of two steps: preprocessing to reduce the number of association candidates and associating objects by using the probabilistic relative distance. We set up an experimental environment including 10 cameras and achieve a better performance than using appearance features only.

**Keywords:** Multiple camera tracking, feature fusion, metric learning.

## 1 Introduction

Maintaining trajectories of objects in a wide area is an interesting research issue for many surveillance systems. If the camera network in the system has a full coverage of the surveillance area, objects probably can be monitored in a wide area. Due to many issues such as cost of cameras, storage and communication bandwidth, it is not affordable to have a full coverage camera network in the surveillance area in most cases. Hence, tracking objects over non-overlapped cameras becomes important for surveillance systems. Although there have been significant improvements in the multiple non-overlapped camera tracking, this problem is still an open issue because of challenges such as ambiguities in appearance of objects, light changing between cameras, and changing of object poses in multiple cameras.

With the recent developments of tracking methods [1], [2], [3], performance of object tracking under occlusions improves significantly. These methods can handle a crowd up to 30-40 objects in a local camera. When objects go from a camera to another one, it is required to have methods to re-identify them and maintain their trajectories. If cameras are overlapped, some methods can be used such as [4], [5]. However, these methods are not suitable for multiple non-overlapped camera tracking due to assumptions on overlapped cameras.

C. Gurrin et al. (Eds.): MMM 2014, Part I, LNCS 8325, pp. 365–374, 2014.
© Springer International Publishing Switzerland 2014

The problem of multiple non-overlapped camera tracking can be summarized as follows: given tracks of objects which disappeared in cameras and tracks of objects just appeared in other cameras, we need to find the correspondences between disappeared tracks and new appeared tracks. Most methods for multiple non-overlapped camera tracking consists of two steps: feature extraction and object association.

Features in multiple non-overlapped camera tracking can be appearance features and spatio-temporal features. First, a brightness transfer function is introduced to compensate the lighting change between cameras [6]. Then, features are extracted for the object association. Appearance features include color features in [7], [8], [9], textures (Gabor features) [9], [10], covariance features [11], histogram of gradient orientation [12] and local descriptor features [13]. Features can also be color name [14]. They can be fused together to improve the performance. Using appearance features can find objects that have consistent appearance across multiple cameras. However, they also find difficulties when other objects are similar in appearance. To overcome this problem, some methods proposed to use the spatio-temporal features [15], [16], [17]. However, these features do not contain enough information when the direction of tracks changes when the objects move between cameras.

For object association, some methods can be applied such as Markov chain Monte Carlo [16], support vector machine (SVM) [9], probabilistic relative distance (PRD) [10], rankboost [14], Munkres assignment algorithms [15]. It is observed that the object association will be better if the number of candidates for the association is reduced.

In this paper, we propose a spatio-temporal feature to take care of changes of directions when objects move between cameras. Appearance features are obtained from color distributions by discriminative color representations [18]. Then, appearance and spatio-temporal features are fused in PRD [10]. Moreover, a preprocessing step is to reduce the number of track candidates by using time and movement directions is also proposed in the paper.

## 2 Problem Description

The problem can be described as follows. Let $\mathcal{T} = \{T_1, T_2, ..., T_N\}$ be tracks of objects moved out of cameras and $\mathcal{S} = \{S_1, S_2, ..., S_L\}$ be tracks just appeared (at least 5 frames) in cameras. $\mathcal{T}$ and $\mathcal{S}$ can be obtained from single camera tracking methods. For each track $S \in \mathcal{S}$, the problem is to find track $T \in \mathcal{T}$ that can be associated with $S$. If track $T$ can be found, the ID of track $S$ will be the same with the ID of track $T$ and some features of track $T$ can be transferred to track $S$. Otherwise, a new ID will be assigned to track $S$. When the number of tracks in $\mathcal{T}$ is large, the performance of the association can be reduced. It is observed that finding a person in a crowd is much more difficult than finding a person in a less crowded situation. Hence, in the next section,

we will propose a preprocessing method to reduce the number of candidates for the object association. Sec. 4 will introduce the feature extraction and the object association method.

## 3 Preprocessing for Object Association

Let $S \in \mathcal{S}$ be a track. When the number of tracks in $\mathcal{T}$ is large, the possible of wrong associations with $S$ will increase. In this section, we try to reduce possible candidates in $\mathcal{T}$ to associate with $S$. Let $T \in \mathcal{T}$ be another track. Track $T$ will be a potential candidate of associating with $S$ when it is satisfied three constraints: camera topology constraint, time constraint and direction constraint.

First, camera topology constraint allows objects move from one camera to defined neighbor cameras. The camera topology is described by graph $G^c = \{V^c, E^c\}$, where $V^c$ is a set of cameras and $E^c$ is a set of edges represented for movements between cameras. Weight $w$ for an edge is 1 if the movement between cameras on this edge is allowed. Otherwise, this edge weight will be 0. The camera topology constraint is defined as

$$\text{TopologyConstraint}\,(S,T) = \begin{cases} 1, \text{ if } w\,(v_T, v_S) = 1 \\ 0, \quad \text{otherwise} \end{cases} \tag{1}$$

where $v_T, v_S$ are cameras capture $S$ and $T$. The time constraint will be

$$\text{TimeConstraint}\,(S,T) = \begin{cases} 1, \text{ if } \alpha_1 < t_b\,(S) - t_e\,(T) < \alpha_2 \\ 0, \quad \text{otherwise} \end{cases} \tag{2}$$

where $t_e\,(\cdot)$ and $t_b\,(\cdot)$ are functions to extract the end time and the begin time of tracks. $\alpha_1$ and $\alpha_2$ are time thresholds for moving between cameras. The direction constraint will be

$$\text{DirectionConstraint}\,(S,T) = \begin{cases} 1, & \text{if } dir\,(T) = D\,(v_T | E^c\,(v_T, v_S)) \\ & \wedge dir\,(S) = D\,(v_S | E^c\,(v_T, v_S)) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $v_T, v_S$ are camera nodes in $G^c$ represented for cameras which capture $T$ and $S$. $dir\,(\cdot)$ is the direction function of a track defined as in Fig. 1. $D\,(\cdot)$ is the direction function of movements between cameras that is trained by experimental data.

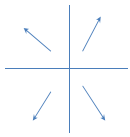

**Fig. 1.** Directions of tracks move between cameras

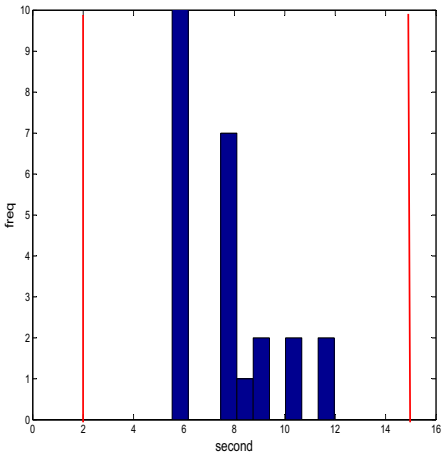


**Fig. 2.** Time thresholds for Time Constraint. Red lines are upper bound and lower bound for time constraint.

## 4 Object Association across Multiple Cameras

### 4.1 Spatio-temporal Feature

Let consider track $S \in \mathcal{S}$ and the track $T \in \mathcal{T}$. The spatio-temporal feature for associating $T$ with $S$ is defined as $f^s(S,T) = \{f_{a_e}, f_{a_s}\}$ where $f_{a_e}, f_{a_s}$ are angle differences between track $T$ and $S$ with typical tracks. They can be obtained as follows. Let $\mathcal{K} = \{\{T_1, S_1\}, \{T_2, S_2\}, ..., \{T_P, S_P\}\}$ be a set of training tracks move from the camera capture $T$ to the camera capture $S$. A set of directions of training tracks is obtained from $\mathcal{K}$, $\mathcal{D} = \{\{d_1^e, d_1^s\}, \{d_2^e, d_2^s\}, ..., \{d_P^e, d_P^s\}\}$ where $d_i^e$ is the direction of $T_i$ and $d_i^s$ is the direction of $S_i$. The direction of a track is the vector from the start position to the end position of the track. After using K-Mean clustering on $\{d_1^e, d_2^e, ..., d_P^e\}$ and $\{d_1^s, d_2^s, ..., d_P^s\}$, we can have typical directions of moving between cameras $D^e = \{\bar{d}_1^e, \bar{d}_2^e, ..., \bar{d}_Q^e\}$ and $D^s = \{\bar{d}_1^s, \bar{d}_2^s, ..., \bar{d}_Q^s\}$ where $Q$ is the number of clusters. An example of extracting representation vectors is shown in Fig. 3. Spatio-temporal feature $f_{a_e}, f_{a_s}$ will be

$$f_{a_e} = \min_{\bar{d}^e \in D^e} \left(d_{\cos}\left(v\left(T\right), \bar{d}^e\right)\right) \tag{4}$$

$$f_{a_s} = \min_{\bar{d}^s \in D^s} \left(d_{\cos}\left(v\left(S\right), \bar{d}^s\right)\right) \tag{5}$$

where $d_{\cos}(\cdot)$ is the cosine similarity distance and $v(\cdot)$ is the function to extract the vector from the start position to end position of a track.

### 4.2 Appearance Feature

In this paper, the color distribution is applied to obtain the appearance feature. Due to illumination changes between cameras, color values of image patches of a

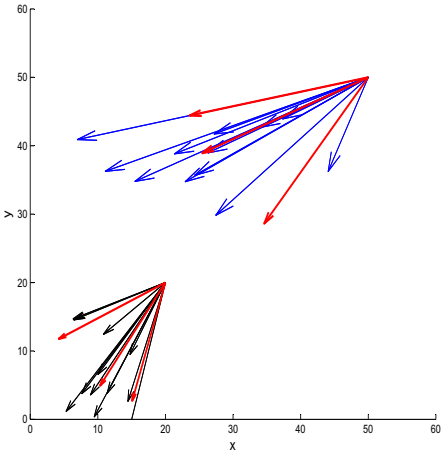


**Fig. 3.** An example of vector representations for movements between two cameras. Red vectors are representations, blue vectors are from the previous camera and black vectors are from the current camera.

person are also not consistent when this person moves across cameras. This cause many difficulties in using color distribution to re-identify persons. Fortunately, in [18], Khan etal. tried to cluster color values based on the Divisible Information Theoretic Clustering method [19]. Color clusters will be trained so that they have an optimum discriminative power of classification for the training data. This discriminative power is measured by the mutual information theory. These clusters are called discriminative color representations. Here, we use 25 color clusters from [18]. The color distribution for the track $S$ is $h(S) = \left[h_1^S, ..., h_{N_H}^S\right]$, where $N_H = 25$, and color distribution for the track $T$ is $h(T) = \left[h_1^T, ..., h_{N_H}^T\right]$. An example of the robustness of the color distribution by using the discriminative color representations is shown in Fig. 4. In this figure, although color changes due to different illuminations, densities of major colors in two color distributions are still similar. The Hellinger distance for these two color distributions is $\frac{1}{2}\sum_{i=1}^{N_H}\left(\sqrt{h_i^S} - \sqrt{h_i^T}\right)^2$. Hence, the appearance feature for associating tracks is defined as

$$f^c(S,T) = \left[\left|\sqrt{h_1^S} - \sqrt{h_1^T}\right|, ..., \left|\sqrt{h_{N_H}^S} - \sqrt{h_{N_H}^T}\right|\right] \tag{6}$$

### 4.3 Probabilistic Relative Distance for Fusing Appearance and Spatio-temporal Features

Features for an association between the track $T \in \mathcal{T}$ and the track $S \in \mathcal{S}$ will be $f(S,T) = \{f^s(S,T), f^c(S,T)\}$, where $f^s(S,T)$ is the spatio-temporal feature in Sec. 4.1 and $f^c(S,T)$ is the appearance feature in Sec. 4.2. For simplification, $f(S,T) = f_T$. Let consider $T, T' \in \mathcal{T}$ where $T$ is the correct match to $S$. In [10],

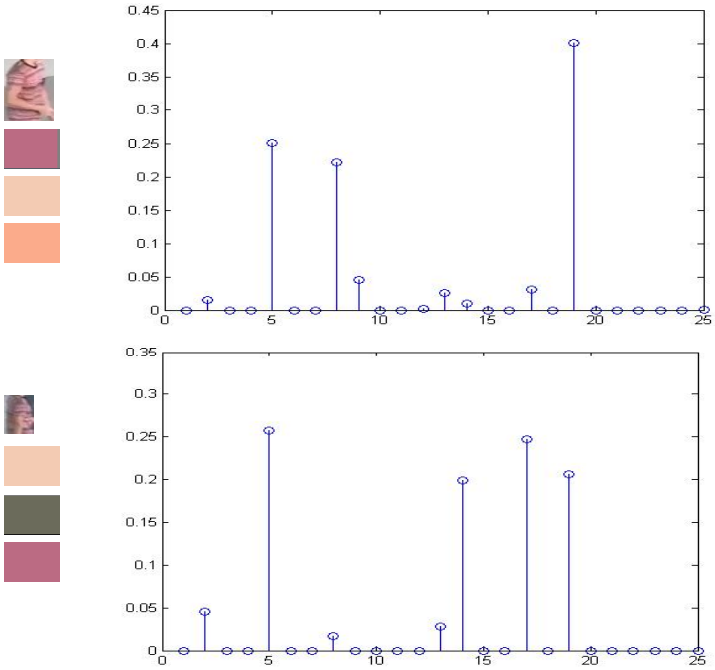


**Fig. 4.** Color distribution by using discriminative color representations when an object acrosses two cameras

a probabilistic relative distance function $r$ is proposed so that $r(f_T) < r(f_{T'})$. $r$ is defined as

$$r(f) = f^T \mathbf{M} f, \quad \mathbf{M} \succeq \mathbf{0} \tag{7}$$

where $\mathbf{M}$ is a semi-definite matrix. $r$ will be trained so that the probability of $r(f_T) < r(f_{T'})$ needs to be maximized. The probability of $r(f_T) < r(f_{T'})$ is

$$p(r(f_T) < r(f_{T'})) = \frac{1}{1 + \exp\{r(f_T) - r(f_{T'})\}} \tag{8}$$

### 4.4 Object Association

For each track $S \in \mathcal{S}$ and track $T \in \mathcal{T}$, the feature for the association will be $f(S,T)$ as in Sec 4.1 and Sec. 4.2. The weight for the association is the probabilistic relative distance as in Sec. 4.3 and [10]. Then, the association problem becomes finding $\hat{\mathcal{X}} = \{X_{ST}, (S \in \mathcal{S}) \wedge (T \in \mathcal{T}), X_{ST} \in \{0,1\}\}$ so that

$$\hat{\mathcal{X}} = \arg\max_{\mathcal{X}} \sum_{S \in \mathcal{S}} \sum_{T \in \mathcal{T}} w(S,T) X_{ST} \tag{9}$$

$$\sum_{S \in \mathcal{S}} X_{ST} = 1, \sum_{T \in \mathcal{T}} X_{ST} = 1$$


**Fig. 5.** Camera coverage of our experimental environment

Note that when $w(S,T) < \eta$, $w(S,T) = 0$. That means the association between $S$ and $T$ is not enough confidence. The problem in Eq. (9) can be solved by using the Munkres algorithm.

## 5 Experimental Results

To evaluate the performance of the proposed method, we set up a camera network including 10 cameras. The coverage of cameras is shown in Fig. 5. Some cameras are different in both zoom and angle, for examples camera 6 and camera 7. The Illumination different is large in some cameras such as camera 3 and camera 4. We collected and annotated 20 videos from 10 cameras at different time. Each video is about 20 minutes. Half data is used for training. For local camera tracking, we applied a method to fuse head detection with single object tracking results for multiple object tracking. The local tracking method can track up to 30 persons per camera view. When a person moves from one camera to another camera and

| Cameras | PRD with color distribution | Our system |
|---|---|---|
| C5-C4 | 68.6% | 88.6% |
| C2-C3 | 63.3% | 83.3% |
| C7-C3 | 57.2% | 78.2% |
| C4-C3 | 53.3% | 71.43% |

**Fig. 6.** Detection rate for person re-identification on four best camera pairs

has well-defined appearance features (not occluded by other persons), we apply the proposed method to re-identify this person. The results on detection rate for camera pairs are shown in Fig. 6. The detection rate is defined as

$$\text{Detection rate} = \frac{\text{number of correct association detections}}{\text{number of ground-truth associations}} \tag{10}$$

The reason that our method is better than the PRD with color distribution is that we have a preprocessing step to reduce the number of association candidates and the fusing between spatio-temporal and appearance features. The overall accuracy of our method is about 71% on this data set. One example of the association is shown in Fig. 8. This example demonstrates the ability of our method when associating objects across multiple cameras given multiple. The number of candidates for association in this case is 59. Our method still can re-identify the person correctly. Some false associations are shown in Fig. 7. False associations can be caused by occlusions and ambiguous in appearance of persons whose have similar movements with training data. To improve the algorithm with these cases is a challenging task.

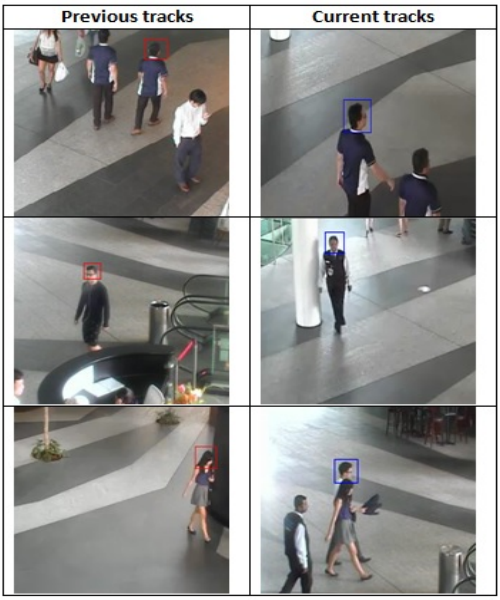


**Fig. 7.** Typical false alarm cases for association


**Fig. 8.** An example of the object association in multiple camera tracking. Red track is a previous track and blue track is a new track.

## 6 Conclusions

In this work, we proposed a spatio-temporal feature and fuse it with appearance features for multiple non-overlapped camera tracking. The method consists of two stages: preprocessing to reduce the number of object association candidates and object association. Our method can be applied for real time surveillance applications. Results also showed that our method is better than using appearance feature only for multiple non-overlapped camera tracking.

## References


1. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: International Conference on Computer Vision (2009)
2. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
3. Andriyenko, A., Schindler, K.: Global optimal multi-target tracking on a hexagonal lattice. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 466–479. Springer, Heidelberg (2010)
4. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-camera people tracking with a probabilistic occupancy map. IEEE Transaction on Pattern Analysis and Machine Intelligence (2008)
5. Eshel, R., Moses, Y.: Tracking in a dense crowd using multiple cameras. International Journal of Computer Vision (2010)
6. Prosser, B., Gong, S., Xiang, T.: Multi-camera matching under illumination change over time. In: Workshop on Multi-Camera and Multi-modal Sensor Fusion Algorithms and Applications (2008)
7. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Computer Vision and Image Understanding (2008)
8. Porikli, F.: Inter-camera color calibration using cross-correlation model function. In: IEEE International Conference on Image Processing (2003)
9. Processer, B., Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by support vector machine. In: British Machine Vision Conference (2010)
10. Zheng, W.S., Gong, S., Xiang, T.: Re-identification by relative distance comparison. IEEE Transaction on Pattern Analysis and Machine Intelligence (2013)
11. Corvee, E., Bak, S., Bremond, F.: People detection and re-identification for multi surveillance cameras. In: International Conference on Computer Vision Theory and Applications (2012)
12. Martinel, N., Micheloni, C.: Re-identify people in wide area camera network. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2012)
13. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristiani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE International Conference on Computer Vision and Pattern Recognition (2010)
14. Kuo, C.H., Khamis, S., Shet, V.: Person re-identification using semantic color names and rankboost. In: IEEE Workshop on Applications of Computer Vision (2013)
15. Kuo, C.-H., Huang, C., Nevatia, R.: Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 383–396. Springer, Heidelberg (2010)
16. Meden, B., Lerasle, F., Sayd, P.: MCMC supervision for people reidentification in nonoverlapping cameras. In: British Machine Vision Conference (2010)
17. Chen, K.W., Lai, C.C., Hung, Y.P., Chen, C.S.: An adaptive learning method for target tracking across multiple cameras. In: IEEE International Conference on Computer Vision and Pattern Recognition (2008)
18. Khan, R., Weijer, J.V.D., Khan, F.S., Muselet, D., Ducottet, C., Barat, C.: Discriminative color descriptors. In: IEEE International Conference on Computer Vision and Pattern Recognition (2013)
19. Dhillon, I., Madella, S., Kumar, R.: A divisive information theoretic feature clustering algorithm for text classification. Journal of Machine Learning Reserach (2003)