Real-Time Multi-Target Multi-Camera Tracking with Spatial-Temporal Information

Xindi Zhang, Ebroul Izquierdo
Multimedia and Vision Research Group, School of Electronic Engineering and Computer Science,
Queen Mary University of London Mile End Road, London, UK
{*xindi.zhang, ebroul.izquierdo*}*@qmul.ac.uk*

***Abstract*—Video security monitoring has always been an important mission for safety reason. In an entire surveillance system, there are usually several cameras distributed sparsely to cover a wide range of public areas (e.g., school, shopping mall or infrastructure). Tracking person through this cameras network is challenging due to different camera perspectives, illumination changes and pose variations. Several algorithms for Multi-Target Multi-Camera tracking (MTMCT) have been proposed in offline method which has delay in getting result. Addressing the need for real-time computation of people tracks through multi camera, the paper proposes an online tracking algorithm. The contributions include (1) online real-time framework which can be used in practical application, (2) extend a single camera multi object tracking (MOT) algorithm to be suitable for multi-camera tracking and (3) use spatial-temporal information to strengthen cross camera person recall performance. The proposed algorithm has been benchmarked against the literature review of MTMC algorithms.**

***Index Terms*—multi-object tracking, multi-camera tracking, video surveillance**

I. Introduction

Intelligent video surveillance for identifying suspects is an active research domain within computer vision. In this regard, the reported algorithms have focussed on the use of single camera multi-object tracking (MOT). Several existing Multi-Target Multi-Camera Tracking (MTMCT) algorithms reported in literature are based on offline method which requires to consider before and after frames to merge tracklets, and do post processing to merge the trajectory. In the literature, hierarchical clustering [1] and correlation clustering [2] are reported for merging the bounding box into tracklets from neighbor frames. In that case, the tracking is hysteresis(delay in outputting final results) which cannot track the person in-time and get the current exact location.

Addressing the need to generate real-time tracker without the apriori knowledge of person tracks, in this paper, an online real-time MTMCT algorithm is developed, which aims to track a person cross camera without overlap through a wide area. The framework do person detection based on Openpose [3]. Build a multi-camera tracker extended by a single camera tracker MOTDT [4]. The novelty of the proposed solution relies on adding a new tracking state. Due to the variation among different perspective of different cameras, the appearance feature is not robust enough to associate person cross camera. To address this issue, the spatial-temporal information are used to mitigate the influence of different views. The main difference is the framework is online and real-time and have competitive performance among other online tracker.

The rest of the paper is organised as follows. Section 2 outlines the literature review within the scope of research presented in paper followed by the proposed algorithms for multi-track person tracking in Section 3. A detailed experienmental evaluation of the proposed algorithm on DukeMTMC dataset is presented in Section 4. The contribution of the paper is summarised in Section 5.

II. Related Work

A. Person Re-Identification

The research on person re-identification has attracted attention from several researchers focused on the development of reliable tracking algorithm. Person Re-ID has been regarded as a classification problem or verification problem. Classification problem uses ID or attributes as labels to train the network, while the verification problem is aim to determine whether the two images belong to one person. The loss function is design to make the distance of the positive pair as small as possible. Common methods are contrastive loss, triplet loss and quadruplet loss [5]. In order to improve the performance, lots of research focus on local feature instead of the global feature of the whole person, such as slice, pose and skeleton alignment [6]. While matching local features help to improve in Person Re-ID, the challenge of pose variation remain open due to the different view from camera.

B. Multi-Object Tracking

Multi-target tracking (MOT) aims to simultaneously locate and track multiple targets of interest in the video, maintain the trajectory and record the ID. Compare to single object tracking, there are two more challenges: the number of targets varies with time, maintain the ID of the targets. MOT algorithms can be broadly classified into two categories namely (i) online and (ii) offline [7]. The online tracking only consider the information of previous and present frame and use current observations to extend existing trajectories gradually. While offline tracking can use future information which can link several observations into trajectories but has a delay in final result output.

978-1-7281-3723-0/19/$31.00 ©2019 IEEE

C. Cross Camera Association

Compare to single camera tracking, multi-camera tracking need to associate the same ID through different cameras without overlapping. For person association, person re-ID features [2] and simple average color histogram [8] are used. In addition to appearance feature, the spatial and temporal information based on the position of cameras also can be considered [9]. Although some of multi-camera tracker have a good performance, they are offline framework which cannot get result in real-time for practical use.

Addressing the influence of pose variation, triplet loss and part-alignment [10] are used to train the feature extraction network by learning to align local parts of interest. In order to build a real-time online framework, the online tracker MOTDT [4] is used to do single camera tracking. We extend it to be a multi-camera tracker. To enhance the performance of multi-camera association by overcoming the current limitation of perspective variation, the spatial-temporal matrix [11], which was used in Re-ID task, are implemented in MTMC tracking task. The details is discribed in next section.

III. Proposed Approach

In this online system, all cameras videos are processed together at the same time frame by frame in multi-thread environment, without post-processing. The proposed algorithm for MTMC includes four stages. In the first stage, person detection is obtained by Openpose [3]. Then, pose points extracted by Openpose are transferred to bounding box coordinates. After refinement, the person feature of each bounding box is extracted and set ID to each of them. In single camera, the tracklet is merged by considering the appearance feature extracted by Re-ID network and motion feature extracted by Kalman Filter. When the ID is disappear in one camera, it will be placed into searching pool and may be reactivated by one of other cameras through its appearance features and spatial-temporal features. The spatial-temporal probability metric is developed by a fast Histogram-Parzen (HP) method. The flow chart of the whole process can be seen in Fig. 1.

A. Person Detection and Reginement

For person detection, we use Openpose which extract points of persons joints . While this points need to be transferred to bounding box coordinates. The detecctor generates a bunch of false positive candidates. So bounding box refinement need to be done by a lightweight RFCN described in [4]. The input of this network is the frame and the bounding boxes. It extracts the feature of the whole frame and does classification on each potential region. The sharing feature map is computational efficiency. After the classification, false positive bounding boxes can be removed.

B. Single Camera Person Association

The tracking algorithm is aimed to merge the bounding boxes of different frames into one track with the same identification. In order to achieve the right combination, the

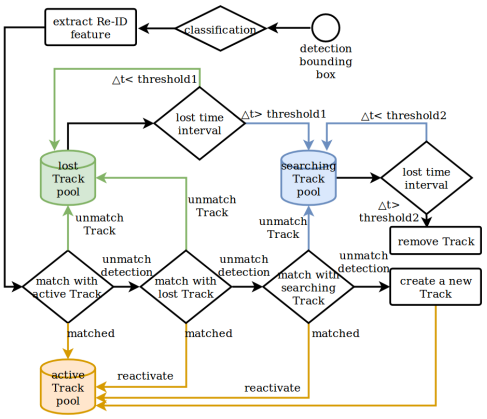

Fig. 1. Flow chart of the framework. In every frame, false positive bounding boxes are reduced. Then extract appearance feature to match with active Track. Combining appearance feature with motion feature to match lost Track and then using spatial-temporal feature with appearance feature to match searching Track. Detection without matching will be create as a new Track. Different states of Track will be placed in different pool, waiting for association with new detection box. And the state will be updated every frame.

appearance features and motion features are used. The appearance features are extracted by the part-align Re-ID network [10] on each bounding box. The backbone of the network H_{reid} is GoogLeNet [12]. It connected to K branches of fully-connected layers for part-alignment. The feature of candidate person I is $f = H_{reid}(I)$. The bounding boxes in the different frames will be merged into one track if the Euclidean distance d_{ij}, between the two candidates I_i and I_j, is smallest among all the distances and within a threshold m. The motion features are generated by Kalman Filter which predict the position of a moving object. The association will be removed if the distance of two bounding boxes is exceeded the predicted area. When a person is occluded by other person or obstacle, Kalman Filter can help to predict the trajectory of the missing target. And when the person reappears, the lost track can be reactivated. When the track reactivate, the Kalman filter will be reinitialized, cause the accuracy will decrease without update over a long time.

C. Cross Camera Person Association

For multi-camera tracking, a person should be correctly associated with the previous $Track$. The appearance feature, spatial and temporal feature is used to do the person association. The appearance feature is extracted by person Re-ID network, and calculate the distance between the new target and the features stored in the $Track$. A spatial-temporal probability metric [11] is used for helping alleviate the problem of appearance ambiguity due to the perspective variation. The spatial-temporal information can be learnt depends on the position of the camera. The time interval between different cameras varies. We summarize the histogram of time interval distribution of possible camera change. And smoothed by the Parzen Window method. The probability of positive associa-

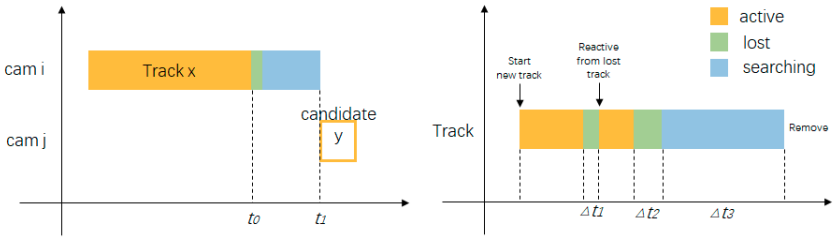

Fig. 2. (a) A Track x in camera i end tracking at time t_0 and get into searching state. A candidate y in camera j at time t_1 will match with Track x depends on the appearance feature and time interval $t_1 - t_0$ related to spatial information (camera transfer from i to j). (b) Each Track has four states. At beginning, a track will be active by create a new track. When the Track lost tracking and if the time interval smaller than the threshold like Δt_1 , it will be reactive. If the time interval larger than the threshold like Δt_2, it will change into searching state. If the time period of searching state is larger than the threshold like Δt_3, the Track will be removed.

tion pair is

$$\hat{p} = (y = 1|k, c_i, c_j) = \frac{n^k_{c_i c_j}}{\sum_l n^l_{c_i c_j}} \tag{1}$$

k means the k-th bin of a histogram. c_i and c_j are the index of camera. $n^k_{c_i c_j}$ represents the number of person pairs disappear from camera i and reappear in camera j in k time intervals. y=1 when the identity I_i and I_j is same. The histogram is smoothed by

$$p(y = 1|k, c_i, c_j) = \frac{1}{Z} \sum_l \hat{p}(y = 1|l, c_i, c_j) K(l - k) \tag{2}$$

$K(\cdot)$ is a gaussian function kernel and $Z = \sum_k p(y = 1|k, c_i, c_j)$ is a normalized factor. Then the appearance feature and spatial-temporal feature are integrate by Logistic Smoothing (LS) similarity metric.

$$P_{joint} = f(s; \lambda_0, \gamma_0) f(p_{st}; \lambda_1, \gamma_1) \tag{3}$$

p_{joint} stand for $p(y = 1|x_i, x_j, k, c_i, c_j)$, p_{st} is $p(y = 1|k, c_i, c_j)$ and s is $s(x_i, x_j)$ which is the similarity score of the appearance feature. $f(\cdot)$ is a logistic function:

$$f(x; \lambda, \gamma) = \frac{1}{1 + \lambda e^{-\gamma x}} \tag{4}$$

so that p_{joint} is robust enough for rare events, since the spatial-temporal probability if not reliable for every situation.

Fig. 3. Histogram of time interval (ID transfer from camera 2 to camera 1). Camera topology of eight cameras.

D. MTMC Tracker

Our tracker is an online tracker which run in real-time without any post-processing. The single-camera multi-object tracking algorithm is [4]. We extend it to be suitable for multi-camera tracking. Each person has his/her own $Track$. Every $Track$ has information, such as Track state, start tracking frame, end tracking frame, which camera is the $Track$ belongs to, and 100 most recent appearance features of the $Track$. There are four different track states: active, lost, searching and removed, as shown in Fig. 2. Active means the person is being tracked in a single camera. Lost means the $Track$ is temporarily lost due to occlusion by other person or obstacles. It will be reactivated soon if the time interval is within the threshold. A $Track$ disappears in a camera will be marked as searching state. This kind of $Track$ will be put into a searching pool. When a new person appears in a camera, it will be matched with the $Track$ in the searching pool depends on the appearance feature and the spatial-temporal feature. A $Track$ disappears longer than a threshold will be marked to be removed which will not be recalled by other cameras.

IV. Experiments

Fig. 4. Experiment example. Person $ID22$ and $ID23$ in $Camera2$ at frame number 7299(on the left), and in $Camera1$ at frame number 9312(on the right). This result shows the correct cross-camera association.

A. Dataset Description

We run the experiment on DukeMTMC dataset [13], which contain 8 cameras with four sequence: trainval, trainval-mini, test-easy and test-hard. The ground truth of testing set is unavailable, so we use the 'trainval-mini sequence' as testing set and the remaining of 'trainval sequence' as training set.

B. Experimental Setup

For appearance feature extraction, the network was trained based on DukeMTMC Re-ID dataset. The k in part-align is 8. The network extract 8 parts-align features inside the bounding box, and concatenate them together as a 512 dimension feature map. To learn spatial-temporal metric, the ground truth of training set are used. For each ID, consider the first frame and the last frame in a certain camera. Then sort the camera according to the frame number. Then calculate the time interval between different camera, and summarize the frequency in every 100 frames to get the histogram (shown in Fig. 3). The parameters for joint metric is: $\lambda_0 = 1$, $\lambda_1 = 2$, $\gamma_0 = 5$ and $\gamma_1 = 5$. The experiment is executed with NVIDIA GeForce GTX 1060 6GB. The processing frame rate at testing stage is 21fps with 8 cameras together. So it achieved real-time and online with high performance. Fig. 4 shows the example.

C. Evaluation Protocol

In order to evaluate the performance, we follow the ID measures of performance in [13]. For ranking MTMC trackers, IDF1 is the principal measure which is the number of correctly identified detection divided by the average number of computed and true detections. IDP, stand for identification precision, and IDR, stand for identification recall, are the scores of true detections which are identified correctly.

D. Result and Discussion

TABLE I
MULTI-CAMERA RESULT IN DIFFERENT SETTINGS

detection	cross-camera association	IDF1	IDP	IDR
DPM	appearance feature	45.41	47.43	43.56
Openpose	appearance feature	47.34	48.95	45.84
Openpose	appearance + ST feature	53.2	55.38	51.96

Table I evaluates the multi-camera result with different configurations. The first two rows indicates the detection bounding box will influence the performance of tracking even after the refinement. The reason is that the refinement network only performs classification without bounding box regression. And the person Re-ID network relay on the coordinate to extract features. DPM generate coarse bounding box around a person with uncertain scale and ratio. The input of appearance feature extraction network will be influenced. While Openpose generate person joints key-points which will perfectly cover a person. So the appearance features are more robust to be matched. The comparison between the second row and the third row shows that the spatial temporal feature help to improve the performance of cross-camera matching. The second row only use appearance feature to do the identity association. The pose and perspective in different cameras are varied. For example, a person in camera 4 is a frontal view. When the person move to camera 3, the view change to side. So the appearance feature are changed that some of person may not be associated. The spatial temporal matrix help to mitigate the influence of pose variation. The combination of appearance feature and spatial-temporal feature increase the match probability of the right identity pair. The performance comparison with other models is shown in Table II. Our models outperform others in IDF1 and IDR.

TABLE II
MULTI-CAMERA RESULT COMPARISON

method	IDF1	IDP	IDR
[13]	37.3	59.6	39.2
[8]	50.1	58.3	43.9
Ours	53.2	55.38	51.96

V. CONCLUSION

In this paper, we developed a multi target multi camera tracking algorithm to track pedestrian in real-time and online. The result proved the detection bounding box will influence the performance of appearance feature matching. And the spatial temporal information help to mitigate the adverse effect of pose variation between different cameras. The limitation still exist for cross camera association, some directions are worth to invest such as camera topology, person gait information in future work.

Acknowledgements. The research activity leading to the publication has been partially funded by the European Union Horizon 2020 research and innovation program under grant agreement No. 787123 (PERSONA RIA project).

REFERENCES

[1] Z. Zhang, J. Wu, X. Zhang, and C. Zhang, "Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project," *CoRR*, vol. abs/1712.09531, 2017.

[2] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," *CoRR*, vol. abs/1803.10859, 2018.

[3] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *CoRR*, vol. abs/1611.08050, 2016.

[4] L. Chen, H. Ai, Z. Zhuang, and C. Shang, "Real-time multiple people tracking with deeply learned candidate selection and person re-identification," *CoRR*, vol. abs/1809.04427, 2018.

[5] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," *CoRR*, vol. abs/1704.01719, 2017.

[6] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," *CoRR*, vol. abs/1701.07732, 2017.

[7] W. Luo, X. Zhao, and T. Kim, "Multiple object tracking: A review," *CoRR*, vol. abs/1409.7618, 2014.

[8] K. Yoon, Y. Song, and M. Jeon, "Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views," *CoRR*, vol. abs/1901.08787, 2019.

[9] X. Chen, K. Huang, and T. Tan, "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognition*, vol. 47, p. 11261137, 03 2014.

[10] L. Zhao, X. Li, J. Wang, and Y. Zhuang, "Deeply-learned part-aligned representations for person re-identification," *CoRR*, vol. abs/1707.07256, 2017.

[11] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," *CoRR*, vol. abs/1812.03282, 2018.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[13] E. Ristani, F. Solera, R. S. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," *CoRR*, vol. abs/1609.01775, 2016.