Multi-Camera Trajectory Mining: Database and Evaluation

Yang Hu, Shengcai Liao, Dong Yi, Zhen Lei, Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences (CASIA)
95 Zhongguancun East Road, 100190, Beijing, China
{yhu, scliao, dong.yi, zlei, szli}@nlpr.ia.ac.cn

***Abstract*—In recent years, large-scale video search and mining has been an active research area. Exploring the trajectory of pedestrian of interest in non-overlapping multi-camera network, namely the *trajectory mining*, is very useful for visual surveillance and criminal investigation. The trajectory mentioned in our work describes the transition of pedestrian among cameras from a macroscopic perspective which is different from the concept in conventional tracking field. In this paper, we collect a database called TMin to promote research and development of trajectory mining. This release of Version 1 contains 1680 images from 30 subjects, all the images are extracted from 6 surveillance videos over two hours, and each subject appears in at least two different cameras. We describe the apparatuses, environments and procedure of the data collection and present baseline performance on the TMin database.**

I. Introduction and Related Work

Recently, there has been increasing research interest in visual surveillance community. Due to the limited field of view (FOV) of a single camera, multiple cameras are needed to construct a camera network covering a range of areas. Intelligent video analysis technology can make use of the video data to maintain public security. In addition, there are relations among large number of videos. Exploring the association of the information of each video can make full use of the captured videos as well as gain more additional information. Some existing work, such as person re-identification and multiple camera tracking, are proposed to achieve the purpose.

Under practical application circumstance, users often want to know the leaving route of subjects or the places they have stayed. Trajectory generated by connecting several key points labeled on a map satisfies the needs well when users care more about the trajectories from a macroscopic perspective rather than the detailed tracking results.

Inspired by the demand described above, we propose a concept called *trajectory mining* which is not the concept in Data Mining field [1]. Our motivation is to explore the trajectory of pedestrian of interest among cameras in non-overlapping multi-camera network. Moreover, trajectory mining is a global optimization problem that concerns about all the cameras at a time which is different from multi-camera tracking who aims to find the correspondences of the observations from two directly connected cameras. The appearance of the same subject may vary dramatically due to different viewpoints, poses and illuminations. In this case, it is difficult to get reliable results if we only consider the appearance information. Therefore, we take temporal information and camera topology into account except for the appearance information to obtain the trajectory. Each FOV of a single camera is regarded as a node, and the trajectory formed by the connecting of several nodes is regarded as the output, it can describe when and where the target had stayed in the multi-camera network.

TABLE I. Contrast between trajectory mining and multi-camera tracking

	Trajectory Mining	Multi-camera Tracking
Topology	Global	Local
Online / Offline	Offline	Online
Input	Images with the time stamps, topology	Sequences obtained by single camera tracking, topology
Output	Trajectory consist of nodes	Correspondences of the sequences

Some existing researches are closely related to trajectory mining such as person re-identification and multi-camera tracking. Compared with person re-identification, trajectory mining needs the time stamps and the topology in addition to image data, however, any of the two information is easy to obtain. In general, trajectory mining is a feasible technology and strongly conforms for the actual demand.

In order to prevent the confusion between trajectory mining and multi-camera tracking, we will further explain trajectory mining from several aspects: (1) Trajectory mining is off-line and aims at processing history videos. Furthermore, all instances with the same identity as the query need to be discovered, regardless of the time lapse and spatial location. In contrast, in multi-camera tracking, if a target disappears for a period of time longer than expected, the target will be assigned a new label [2]. (2) Multi-camera tracking pays more attention on the local structure of the topology, and mainly seeks for the correct correspondences between observations of two directly connected cameras. Some researchers evaluate the tracking performance by calculating the matching accuracies of pairwise cameras [3]. As for trajectory mining, the whole topology will be considered when we seek for the optimal solution and evaluate the performance. (3) The output of trajectory mining is the transition of pedestrian among the nodes rather than the conventional tracking result. We summarize the differences between trajectory mining and non-overlapping multi-camera tracking in Tab.I.

Plenty of work of other researches are helpful to trajectory mining. Person re-identification can match person from

Stan Z. Li is the corresponding author

different camera views which is an essential part of trajectory mining. Some person re-identification methods aim to extract visual features which are both distinctive and stable under various conditions [4–7]. After feature extraction, an established distance measurement is applied to compare different person representations. Some other methods aim to learn the optimal distance measure for many features jointly via distance learning theory [8–10]. These methods are less sensitive to feature selection, therefore they usually use very simple features. There are also plenty of worthy work in multi-camera field that inspire us a lot. Thang and Worring et al. [11] integrated motion detection, object classification, object modeling and matching, interactive retrieval and visualization into a complete working system to track the reoccurrences of objects in multi-camera visual surveillance. In [12], the author solved the tracking problem based on systematically building the link between cameras, the camera link model can be learned based on a fully unsupervised scheme without manually labeling the correspondences in advance [13]. Makris et al. [14] proposed a method to build the transition time distribution based on the cross-correlation function between the exit and entry time stamps of the observations. The brightness transfer function (BTF) [15, 16], which stands for the mapping of color models between two cameras, can be applied to compensate for the color difference between two cameras before we compute the distance between images.

To promote the development of trajectory mining research, we collect a database, called TMin in this paper. Six cameras construct a non-overlapping multi-camera network to cover an area. This release of Version 1 contains a total of 1680 images from 30 subjects. The collection of this database is inspired by some existing person re-identification database such as VIPeR [17] and ETHZ [18], and the TMin database can also be used for person re-identification. However, in non-overlapping multi-camera tracking field, no appropriate public database can be found for our task. Trajectory mining has a requirement for the number of cameras and subjects to generate various trajectories, therefore, the collection of TMin database is meaningful.

In the rest of the paper, Section 2 introduces the structure of the multi-camera network and devices used to collect the database, the data acquisition procedure will be introduced in this section as well. Section 3 describes the content of the database. The baseline performance is presented in Section 4. Section 5 summarizes the paper.

II. Data Acquisition Setup

This section describes the multi-camera network scenarios and the procedure of the video acquisition, pedestrian image extraction methods are introduced as well.

A. Multi-camera Network Topology and Video Acquisition

The multi-camera network consists of six cameras, and the layout of the site is shown in Fig.1. We select six most frequently pass by non-overlapping FOVs as the topology nodes, each node corresponds to a specified camera, therefore they share a same ID. Most trajectories can be described by the connecting lines of these six nodes. We can comprehend the multi-camera network more clearly through the corresponding topology which is shown in Fig.2. If there exists a path allowing people to travel between two cameras without passing through any other cameras, we consider the two cameras directly connected and there will be a link between the two nodes (cameras).

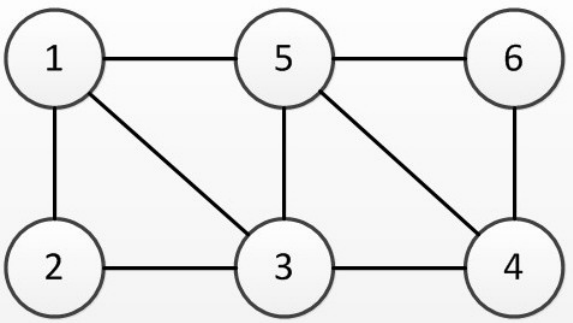

Fig. 2. The topology of the non-overlapping multi-camera network.

We utilize three kinds of camera to collect the video. In the areas of Cam1, Cam2, Cam4 and Cam5, videos are collected by HIKVISION DS-2CD864FWD-E IP camera, for the area of Cam3, we utilize SAMSUNG SNB-6004P IP camera and SAMSUNG SNB-7000P IP camera for the Cam6 area, respectively. Considering all the factors, the positions of the six cameras mounted are marked in Fig.1. We set the image resolution to 1280×720, and the frame rate is 25fps for all the six cameras. In order to capture clear images of pedestrians with more details, we set the camera focal length to the largest and use auto iris to adjust the illumination changes. It is cloudy when recording the videos, so that the overexposure is rare while some videos are a little underexposure. The video acquisition time starts at twenty to twelve in the morning and ends at a quarter to two in the afternoon. During this period of time, more pedestrians can be collected, however, occlusions may make the image extraction of pedestrians more difficult at the same time. Finally, we obtain a 125-minutes video per camera.

B. Pedestrian Images Acquisition

After we get six surveillance videos, the next step is to extract observations of the pedestrians of each video. Due to the high resolution of the videos, we apply background substraction method firstly to improve the calculation speed. In this paper, we apply the background substraction method proposed in [19] which was good to deal with complex dynamic scenes. Then, we use pedestrian detection technology based on deformable part model proposed in [20] to initialize the tracking. Finally, we apply the nearest neighbor tracking strategy and then save the tracking results.

Large amounts of images will be generated after the above three steps and we need to pick out the useful ones. Firstly, we select the persons that have appeared in at least two different cameras. Secondly, we collect all the appropriate images belong to the selected person. Thirdly, we label all the images of the same person a consistent ID. In addition, some other useful information is also labeled manually such as the source of the person (camera ID), the frame and observation ID.

Fig. 1. The layout of the site. The white ellipse areas indicate the FOV of a single camera, and the arrow points to the corresponding actual scenario. The red trapezoid blocks denote the positions where we locate the cameras.**Best viewed in color**.

III. Database Description

The TMin database consists of 30 subjects collected from six cameras, for each subject, there are at least one observation from the camera which the person has appeared in. If a person goes out of the FOV of a certain camera and comes back after a while, two observations will be obtained with the same person ID (the observation IDs are different). The dataset can be downloaded form http://www.cbsr.ia.ac.cn/english/TMin-1.0-Database.html

This release of TMin database includes the following:

1. There are six folders named by the camera ID, each folder contains the images of the pedestrians of interest belong to the corresponding camera. There are a fixed 20 images for each pedestrian in each folder (camera). If a pedestrian have more than one observation in a folder, we will decrease the sampling rate for each observation to make sure the total number of images does not change. The numbers of the pedestrian, observation and image in each folder are shown in Tab.II.

2. The name of each image has three components: global ID, camera ID and frame. The format of the name is: "GlobalID_CameraID_#Frame". All the images are in the JPEG format, we extend the height and width of images 1.5 times the size of the height and width of the bounding box obtained by tracking.

3. For each folder, we provide a text document to record the corresponding coordinates of the pedestrian images in the captured video. In each row of the document, we record the coordinates by appending the x-coordinate of the center, y-coordinate of the center, width and height of the bounding box at the end of the image name, the observation ID is appended at the end of coordinates. There are a total of 94 observations in this release.

4. We also provide the ground truth of the trajectories for these 30 subjects.

Some samples are shown in Fig.3 and Fig.4. Fig.3 shows

Fig. 3. We select three persons from the TMin database as examples and plot their trajectories upon the multi-camera network layout.

the trajectories of several pedestrians. Fig.4 shows some image samples in our database.

IV. Baseline Method

In this section, we evaluate the baseline performance of the TMin database. The color information and the temporal information is used to mine the trajectory.

A. Color Information

The same subject may appear differently under two cameras with non-overlapping views due to occlusion and different poses. Therefore, we apply the method proposed in [5] which aims to solve the person re-identification problem. Considering the calculation speed, we only apply the HWH (hierarchical weighted histograms) which is a weighted color histogram taken into account the structural information of body parts.

Here we review the procedure of extracting HWH: (1) Obtain three basic components by dividing the body sketch into head, torso and leg. (2) Construct a hierarchical structure with 3 layers consist of six partitions. (3) For each partition, a Gaussian kernel is applied to assign different weights when extracting the HSV color histograms. HWH is obtained by concatenated all the six histograms. By assigning higher weight to the central pixels and considering structure information of human body, HWH can solve the pose variation and occlusion to a large degree.

For observation i and observation j, the maximum similarity of two images between the two sets is regarded as the the similarity of two observations. We apply the Bhattacharyya distance to compare the HWH feature vectors f_i and f_j, therefore, the color similarity is obtained as:

$$S_{color}(i,j) = 1 - d(f_i, f_j) \tag{1}$$

B. Temporal Information

To our knowledge, people tend to follow similar paths in most cases due to the presence of available pathways or shortest routes in a specific place. Thus, the transition time forms a certain distribution function. In this paper, we obtain the model of multi-camera network by training the transition time between arbitrary two directly connected cameras and then get 9 distribution function. In addition, we assume the transition time obeys a gaussian distribution. Then, if we have two observations from two directly connected cameras, the likelihood value $S_{tran}(.)$ is considered as the temporal similarity.

$$S_{tran}(i,j) = \exp[-\frac{((t_i - t_j) - \mu_k)^2}{2\sigma_k^2}] \tag{2}$$

where i and j are observation IDs, t_i and t_j are the average frames of the observations, $k = 1, 2, ..., 9$ and k represents the k-th distribution function while μ_k and σ_k^2 are the mean value and variance, respectively.

C. Fusion

Finally, the similarity of two observations is conducted by fusing the appearance similarity and the temporal similarity:

$$S(i,j) = \lambda S_{color}(i,j) + (1-\lambda) S_{tran}(i,j) \tag{3}$$

where the weighted coefficient λ is set 0.5 in this work.

D. Evaluation Setting and Performance

In this paper, we will evaluate the baseline performance on the TMin database as follows:

(1) We randomly select 24 pedestrians for training, then 9 gaussian distribution function correspond to the 9 pair directly connected cameras (see Fig.2) can be obtained.

(2) The left 6 pedestrians are used for testing and we can obtain an observation set $\{O_k\}$ consist of all the observations

Fig. 4. Image samples of the TMin database. Images with a highlighted same color box are from the same person, we only label 3 pedestrian in this figure.

of these 6 pedestrians. We calculate the similarity of arbitrary two observations by Eq.3.

(3) As we know which camera an observation belongs to, we get one observation from each camera, a trajectory consists of six observations is established. The score of the trajectory is the sum of the similarities between the query observation and the other observations selected. After we get all the possible trajectories, we sort the scores in descending order and we reserve the first ten trajectories.

(4) For each trajectory, we calculate the average time of each observation and reorder the trajectory in time order.

(5) Besides, there are two thresholds T_{color} and T_{tem}, if the color similarity is lesser than T_{color} or the temporal similarity is lesser than T_{tem}, the observation will be eliminated.

When the output trajectory contains all the camera IDs of the ground truth trajectory and order is correct as well, we consider the trajectory to be correct. We repeat the above steps for 5 times, and the average top ranked matching rate is used to evaluate the performance. Our recognition rate is 56.25% at rank-10. We have the following conclusions for the baseline performance: The baseline performance of the TMin database is not so good, the chromatic aberration between cameras makes the color similarity unreliable, lack of training samples makes it a difficult problem when a pedestrian goes there and back or has a unusual trajectory. In summary, development of effective methods are needed to achieve high performance.

V. Conclusion

In this paper, we introduce a new concept called trajectory mining and collect the TMin database for trajectory mining research. The imaging apparatuses, environment, acquisition and process procedures are described. Experiment designed for the baseline performance on the database is presented. Although the baseline performance is not good enough, it is a

TABLE II. NUMBERS OF THE PEDESTRIAN, OBSERVATION AND IMAGE IN EACH FOLDER

Folder name	Cam1	Cam2	Cam3	Cam4	Cam5	Cam6
Number of Pedestrian	15	7	20	20	14	8
Number of Image	300	140	400	400	280	160
Number of Observation	17	7	23	23	15	9

way to solve the proposed problem and can help us understand the concept better. In the future, we will enlarge the amount of subjects and apply more effective methods to achieve better performance on the TMin database, such as modeling the BTF between cameras, learning more about the topology and better objective function, some of the work is already in progress.

ACKNOWLEDGMENT

This work was supported by the Chinese National Natural Science Foundation Projects #61105023, #61103156, #61105037, #61203267, #61375037, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, Jiangsu Science and Technology Support Program Project #BE2012627, and AuthenMetric R&D Funds.

REFERENCES

[1] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 330–339. 1

[2] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 952–957. 1

[3] C.-H. Kuo, C. Huang, and R. Nevatia, "Inter-camera association of multi-target tracks by on-line learned appearance affinity models," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 383–396. 1

[4] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *CVPR*, 2010, pp. 2360–2367. 2

[5] Y. Hu, S. Liao, Z. Lei, D. Yi, and S. Z. Li, "Exploring structural information and fusing multiple features for person re-identification." 4

[6] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidentification using spatiotemporal appearance," in *CVPR (2)*, 2006, pp. 1528–1535.

[7] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu, "Shape and appearance context modeling," in *ICCV*, 2007, pp. 1–8. 2

[8] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by support vector ranking," in *BMVC*, 2010, pp. 1–11. 2

[9] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *CVPR*, 2011, pp. 649–656.

[10] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295. 2

[11] T. V. Pham, M. Worring, and A. W. Smeulders, "A multi-camera visual surveillance system for tracking of reoccurrences of people," in *Distributed Smart Cameras, 2007. ICDSC'07. First ACM/IEEE International Conference on*. IEEE, 2007, pp. 164–169. 2

[12] C.-T. Chu, J.-N. Hwang, J.-Y. Yu, and K.-Z. Lee, "Tracking across nonoverlapping cameras based on the unsupervised learning of camera link models," in *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*. IEEE, 2012, pp. 1–6. 2

[13] C.-T. Chu, J.-N. Hwang, Y.-Y. Chen, and S.-Z. Wang, "Camera link model estimation in a distributed camera network based on the deterministic annealing and the barrier method," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 997–1000. 2

[14] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2. IEEE, 2004, pp. II–205. 2

[15] T. D'Orazio, P. Mazzeo, and P. Spagnolo, "Color brightness transfer function evaluation for non overlapping multi camera tracking," in *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*. IEEE, 2009, pp. 1–6. 2

[16] B. Prosser, S. Gong, and T. Xiang, "Multi-camera matching using bi-directional cumulative brightness transfer functions." in *BMVC*, vol. 8, 2008, pp. 164–1. 2

[17] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, vol. 3, 2007, p. 5. 2

[18] A. Ess, B. Leibe, and L. J. V. Gool, "Depth and appearance for mobile scene analysis," in *ICCV*, 2007, pp. 1–8. 2

[19] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikainen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1301–1306. 2

[20] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Multi-pedestrian detection in crowded scenes: A global view," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3124–3129. 2