Coarse-to-Fine Multi-Camera Network Topology Estimation

Chang Xing, Sichen Bai, Yi Zhou, Zhong Zhou*, and Wei Wu*

State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, Beijing, 100191, China
zz@buaa.edu.cn

Abstract. In multiple camera networks, the correlation of multiple cameras can provide us with a richer information than a single camera. In order to make full use of the association information between multiple cameras. We propose a novel approach to estimate a camera topology relationship in a multi-camera surveillance network, which is unsupervised and gradually refined from coarse to fine. First, an improved cross-correlation function is used to get a preliminary result, then a time constraint feature matching model is used to reduce the error caused by external environment and noise, which can increase the accuracy of our results. Finally, we test the proposed method on several different datasets, and its result indicates that our approach perform well on recovering the topology of the camera and can improve the accuracy on over camera tracking.

Keywords: Multi-Camera, Camera Topology, Cross-Correlation, Feature Matching.

1 Introduction

As an important part of the security technology system, video surveillance system is always a good studied topic and research hotspot. The need of practical application on this kind of system has being rising rapidly. However, nowadays, most of the traditional surveillance systems rely on human cooperation and brings burden for operators. For example, if a car needs to be found in a city area, the operator has to search every surveillance video in the area by orders. Even he can find the right car, the operator spend too much time to understand the order relationship of occurrence during the different cameras. It is often ineffective and obtain inaccurate result. Therefore, many intelligent surveillance systems have been presented to inference the relationship between multiple videos automatically.

Unlike single camera, a multi-camera surveillance network has a wider field of view. So, it is hard for us to associate cameras at different positions. Over the past few decades, camera topology relationship is presented to determine the

* Corresponding author

relationship between the cameras. Makris et al. [1] proposed a method to estimate the camera topology relationship in a camera network based on the simple occurrence correlation between entering and exiting events. Kinh Tieu [2] presented an approach to estimate the topology of a camera network by measuring the degree and nature of statistical dependence between observations in different cameras. Unlike previous work, Kinh explicitly considered the correspondence problem and handles general types of statistical dependence by using mutual information and non-parametric density estimates. Niu [3]proposed a model constructed by the combination of normalized color and overall model size to measure the moving objects appearance similarity across the non-overlapping views, their method combines appearance information and statistics information of the observed trajectories, which can overcome the disadvantages of the approaches that only use one of them. Then, based on Kinhs method, Zehavit [4] proposed a method which divides the camera frame into blocks, and refines the relationship between camera into block level. Chen [5] focuses on decreasing the large variance of transition time of true correspondences, which can compensate for the influence caused by large-scale false correspondences to a certain degree.

The methods mentioned above mainly dependent on the relevance of time, and do not take into account the target speed of movement. Their method generally relies on long term videos to reduce the error. To solve those problem, we propose a novel multi-camera network topology estimation method in this paper.

The main contributions of this paper are concluded as follows:

- A coarse-to-fine framework to estimate an accuracy camera topology in the multi-camera surveillance system.
- The proposed approach has good scalability, which can be applied in various field such as over camera tracking which can improve the accuracy and efficiency of existing methods.

2 Our Approach

Our coarse-to-fine multi-camera network topology estimation approach is divided into two main procedures. Firstly, given input entries in videos, we use cross-correlation function model to calculate the transition time and improve it with neighbor accumulate method and peak detection. Then, a time constraint based on feature matching method is used to get a more accuracy transition time distribution. Through the above steps, we can obtain accurate correlation between cameras.

2.1 Improved Cross-Correlation Function Model

As the input of our method, entries in the videos are detected by clustering foreground where objects moving into or leaving from the camera. We consider that objects are directly corresponding to entry. So we uses Faster-RCNN [6],

a constructive work of recent years, to detect the (such as human) location of objects in the camera view. Then the entry zones can been easily clustered by the K-means method [7]. Three examples are shown in Fig.1.

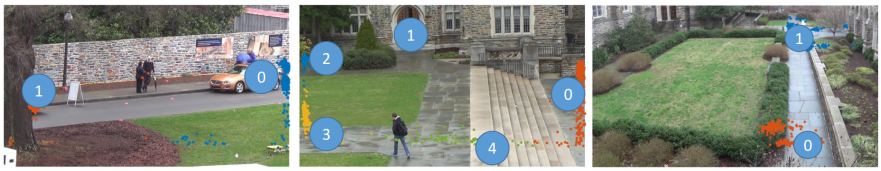

Fig. 1. A example of entries division in cameras (on DukeMTMC[8]).

According to moving direction, the objects are added into corresponding departure sequence and arrival sequence. As for the entry i in camera c_1 and entry j in camera c_2 , if there is a strong transition relationship between them, the object leaves from i at time t_1 will move into entry j at time $t_2, t_2 \in [t_1 - \tau_0, t_1 + \tau_0]$, where τ_0 is a parameter that defines the transition time window. We assume that the transition time obeys the Gaussian distribution. The origin cross-correlation function is defined as $R_0^{ij}(\tau) = E[D_i(t) \cdot A_j(t + \tau_0)]$, where $D_i(t)$ is the departure time sequence at entry i, $A_j(t + \tau_0)$ is the arrival time sequence at entry j. τ_0 is the transition time. $R_0^{ij}(\tau)$ will have obvious peak which represent that there is a transition relationship between the two entries.

However, the origin cross-correlation function uses only temporal information, the state of the cross-correlation function is unstable. In order to get a clear and steady peak, an improved cross-correlation function is introduced to calculate the transition time window. We use an n-neighbor accumulated method [5] to improve the stability of the cross-correlation function:

$$\begin{aligned} R^{ij}(\tau) &= \sum_{\tau_0=\tau_n-5}^{\tau_n+5} R_0^{ij}(\tau_0) \\ &= \sum_{\tau_0=\tau_n-5}^{\tau_n+5} E[D_i(t) \cdot A_j(t + \tau_0)] \\ &= \sum_{\tau_0=\tau_n-5}^{\tau_n+5} \sum_{t=-\infty}^{+\infty} D_i(t) \cdot A_j(t + \tau_0), \tau_n \geqslant 5 \end{aligned} \tag{1}$$

The n is set to 5 empirically, which can solve the problem of excessive accumulation in [5].

Intuitively, at different entries, only those objects which look similar in appearance can be counted to derive the spatio-temporal relation. The $E(,)$ can be transformed to Eqs(2):

$$E\left[D_i\left(t\right)\cdot A_j\left(t+\tau_0\right)\right]=\sum_{O_i\in D_i(t)}\sum_{O_j\in A_j(t+\tau_0)}similiarity\left(O_i,O_j\right) \tag{2}$$

$similiarity\left(,\right)$ denotes the similarity between two objects (O_i, O_j) in corresponding sequences.

A threshold is empirically set to detect the peak interval of the $R^{ij}\left(\tau\right)$ from the mean and variance of the transition time.

$$threshold=avg\left(R^{ij}\left(\tau\right)\right)+\omega\cdot std\left(R^{ij}\left(\tau\right)\right) \tag{3}$$

The value below the threshold is considered to be the noise. After that, we search for a peak interval in the $R^{ij}\left(\tau\right)$, t_k is identified as a candidate if it satisfies the formula $R^{ij}\left(t_k-1\right)\leqslant R^{ij}\left(t_k\right)\leqslant R^{ij}\left(t_k+1\right)$, $W^{ij}\left(t_k\right)$ represent the interval width of $t_k$which is extended until $threshold > R^{ij}\left(t_p\right)$ or $R^{ij}\left(t_p\right) > R^{ij}\left(t_k\right)$. In this work, we assume there is only one popular transition time if there is a link between entry i and entry j. If there is more than one candidate t_m and t_n, a threshold α is set to merge them (α is the width of the candidate, empirically set to 0.2W).

$$W^{ij}\left(t_m\right)=W^{ij}\left(t_m\right)+W^{ij}\left(t_n\right),\qquad if\quad t_m-t_n<\alpha \tag{4}$$

Through the repeated iteration, t_k is the final transition time when there is only one interval. And the transition time window W_0^{ij} approximates to its interval width:

$$T^{ij}=t_k, W_0^{ij}=W^{ij}\left(t_k\right) \tag{5}$$

Otherwise, there is no direct correlation between two entries. Then, a coarse results is obtained. Fig.2(a) is the cross-correlation function without any process, though the accumulation and the peak detection, the peak is much more clear and smooth and its much easier to be recognized (Fig.2(b)).

2.2 Time Constraint Feature Matching Model

The improved cross-correlation function helps us to get a preliminary transition relationship between entries, but it still has a possibility of making error: the speed of the object and some noise such as wrong detection are not taken into account that will make the result unreliable and imprecise. For example, there are three sequences: sequence1 is (0, 0, 0, 1, 2, 3, 4, 5, 6.), sequence2 is (1, 2, 3, 4, 5, 6.). sequence 3 is (1, 2, 4, 3, 5, 6.). When calculate the correlation relationship between sequence1 and sequence2, the transition time is 3 and the cross-correlation function has a clear peak. Due of the reverse of the number 3 and number 4, the transition time between sequence1 and sequence3 is 0.

To solve these problem, we proposed a time constraint feature matching model. First, domain guided dropout algorithm [9] is used to extract the appearance feature of the object in departure sequence and arrival sequence. For entry i and entry j, which already get the transition time T^{ij} and the width W_0^{ij} preliminary

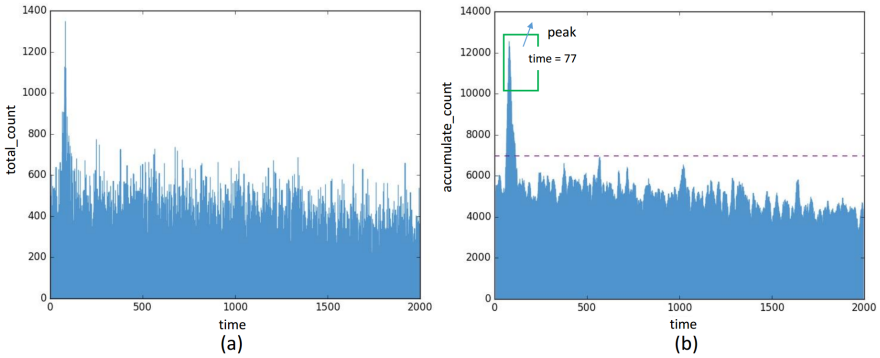

Fig. 2. The coarse result by the proposed method. The dotted line is the threshold. (camera2_zone0, camera1_zone1, on DukeMTMC[8]) (a)Cross-correlation function result without accumulate. (b) Peak detection result.

by the improved cross-correlation function approach mentioned before, T^{ij} follows a normal distribution $X\left(T\right) \backsim N\left(\mu, \sigma^2\right)$ .When an object leaves from entry i at time t, search for the most similar object in the objects sequence moving into entry j during the time transition window $\left[t + T^{ij} - 3 * W_0^{ij}, t + T^{ij} + 3 * W_0^{ij}\right]$, as shown in Fig.3(a). Since the coarse result is already got, it should be a great probability to match the same object in the two sequences. Each matching pair will have a time interval η_0 between them. To calculate the mean and variance for the time interval function $T_0^{ij}\left(\eta_0\right)$, the neighbor accumulated method is used:

$$T^{ij}\left(\eta\right) = \sum_{\eta_0=\eta_n-5}^{\eta_n+5} T_0^{ij}\left(\eta_0\right), \quad \eta_n \geq 5 \tag{6}$$

$T_p^{ij}\left(\eta\right)$ corresponding to the accumulated time interval function. The value is still 1 after accumulated is considered as noise and will be eliminated.

$$\begin{aligned} &T_p^{ij}\left(\eta\right) = T^{ij}\left(\eta\right), \\ &s.t. \eta \in \left\{\eta^{'} | T^{ij}\left(\eta^{'}\right) > 1\right\} \end{aligned} \tag{7}$$

The process of our method is shown in Fig.3. The Fig.3(b) represents the cross-correlation function without any process. Fig.3(c) is the result by using the improved cross-correlation function. The Fig.3(d) shows the time interval $T_0^{ij}\left(\eta_0\right)$ before accumulate, the Fig.3(e) shows that $T_p^{ij}\left(\eta\right)$ is well fitted to the Gaussian function model after processed. The transition time window is easy to get from this figure. By using the time constraint feature matching method, our results are more accurate than before.

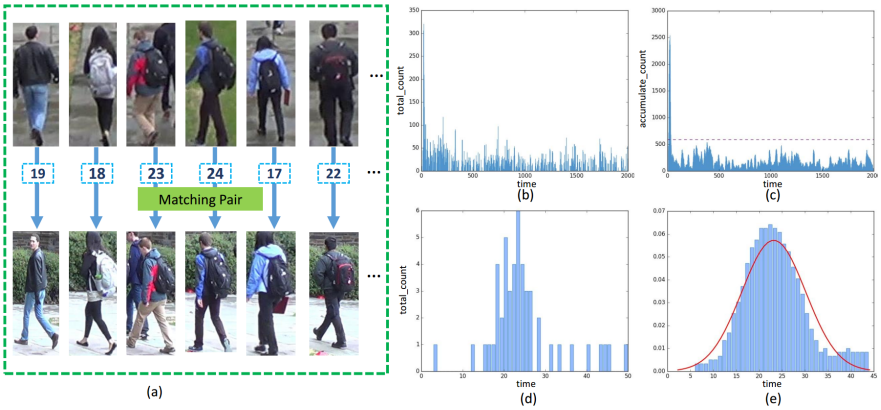

Fig. 3. A example of estimated transition distributions by different method. (a) Feature matching result. (b) Cross-correlation function. (c) Method proposed in section 4.1. (d) Time interval before accumulation. (e) Processed time interval function (on DukeMTMC, camera5_zone0, camera3_zone0).

3 Experiments

To evaluate the effect of the proposed method, we test our method on public datasets: DukeMTMC [8] and NLPR_MCT [5] (including two different scenes: NLPR_MCT_1 includes street and indoor scene, NLPR_MCT_2 is campus monitoring video), which are time synchronously and applicable to the proposed approach. We conduct multiple experiments to give a performance test of the proposed method. After using color transfer as a preprocessing, we test our method on camera topology recovery time and over camera tracking across non-overlapping experiments.

Data Preprocessing. Actually, due to the difference in both lighting conditions and camera parameters, the same object in different camera would have completely different hues, which will result in mistakes on cross-correlation function and lead to failure of feature matching. To make them have consistency in color style, we use a normalization appearance feature model to transfer the color from target camera view to source camera view. The color transfer consists of two parts: the transfer for the luminance and the transfer for the chrominance, which is proposed in our previous work [12]. We use this method to transfer the color style of the object in departure sequences and arrival sequences. As shown in Fig.4, the color style of objects in different cameras turn to be the same. The luminance distribution of the target and source images become consistent. Through this method, the accuracy of the object matching has been remarkably improved (the accuracy of finding the same object in the corresponding sequences increases from 31.4% to 35.0%, on DukeMTMC, camera3_zone1, camera4_zone0).

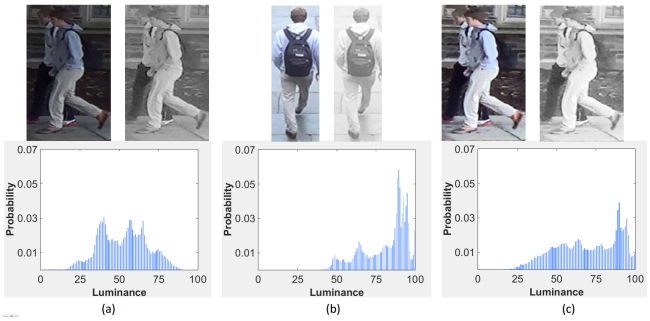

Fig. 4. Color transfer result and luminance cumulative histogram.

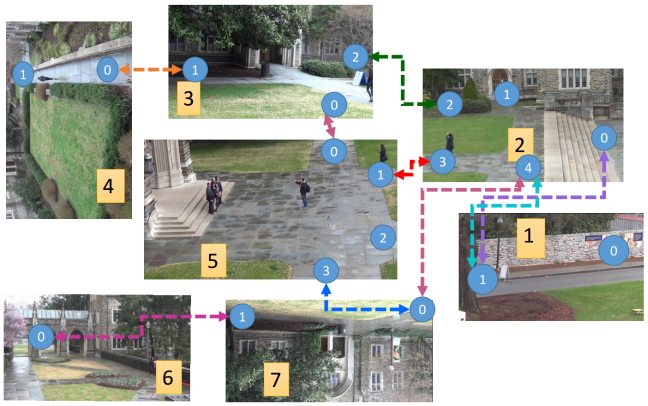

Fig. 5. The camera topology graph by our approach (on DukeMTMC).

Camera Topology Recovery. We estimate the performance of the proposed method on recovering the topology structure and calculate the transition time between entries, we use our method to detect the links between the entries in cameras. Table.1 summarizes the results of the inter-camera correspondence for all the cameras and zones in the camera network on DukeMTMC[8] (the camera topology is shown in Fig.5). μ and σ is expectation and variance of the transition time. In order to simplify the calculation, we extract one key frame for every 20 frames (original video@ 59.940059 fps). All the link between cameras are detected and consistent with the ground truth.

We also compare our approach with previous method. As shown in Fig.6. In Makris's [1] method, the peak is not obvious and difficult to adapt to complex scenes. In Chen's [5] method, a lot of manual parameters is needed and it's difficult to adjust these parameter, furthermore the cross-correlation function is continuously accumulated and excessive accumulation can cause small transition times missing. The previous methods show unclear and noisy distributions for

Table 1. Transition time between cameras

Departure Zone	Arrival Zone	μ	σ
C5,Z0	C3,Z0	19.9	5.6
C5,Z1	C2,Z3	35.2	6.5
C5,Z3	C7,Z0	25.2	5.2
C3,Z1	C4,Z0	54.8	10.0
C3,Z2	C2,Z2	26.2	5.7
C2,Z0	C1,Z1	77.5	14.2
C7,Z1	C6,Z0	24.1	7.4
C1,Z1	C2,Z4	159.8	13.7
C2,Z4	C7,Z0	97.3	8.1

both valid and invalid. As illustrate in Fig.6(c). The transition time by our approach is much more accuracy and does not need extensive tuning experience.

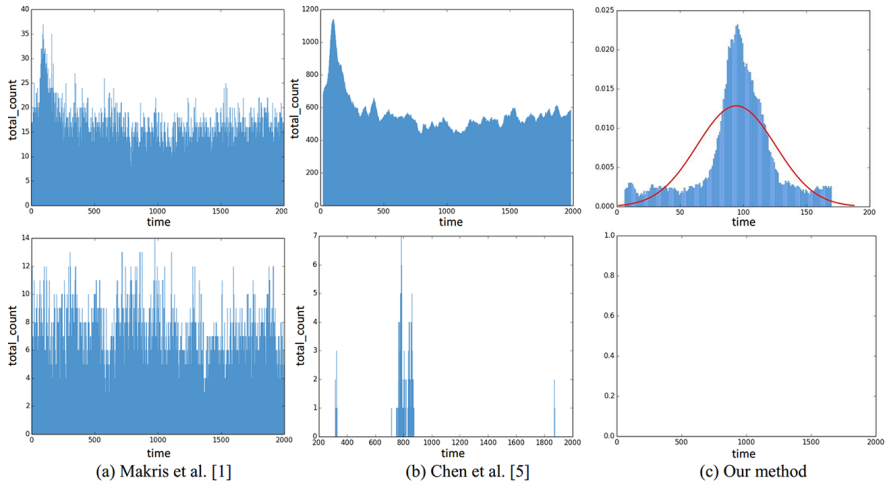

Fig. 6. Performance comparison with other methods. First row: Valid link (camera1_zone1, camera2_zone0). Second row: Invalid link (camera3_zone1, camera5_zone3)

We also validate our approach on other datasets. The result is illustrated in Table.2. The association number represents the count of all the certain link between entries (over camera). The correct detection represents the number of current link number detected by our approach. We have applied the proposed method for each pair of entries. As the result shows, our method performs generally well on various kind of scene but have some error detection. The reason is that although there is a real path between some entries, it fails to be detected as there are too few object moving between these two entries, and there might

Table 2. Result on public datasets.

Dataset \ param	camera	entry zone number	association number	correct detection
DukeMTMC	7	20	18	18
NLPR_MCT_1	3	9	4	4
NLPR_MCT_2	5	11	10	6

have a fork in the blind area between cameras that the crowed is too disperse to have a strong correlation.

Over Camera Tracking. We notices that building topological relationships on multiple cameras can help us to correlate targets in different cameras. The transition time between cameras can also help us on over camera tracking between cameras with disjoint views. We compare the accuracy between using the transition distributions information and not using. The result is illustrate in Table.3. By using our approach, when the target object departures from a camera, we are no longer need to search all the cameras. The highly reliable transition distributions information can help us to find the neighboring cameras. Clearly, this method narrows the retrieval scope and plays a key role in finding out the object accurately and effectively. Meanwhile, the time of matching the independent object is shortened. The data in the table represents the accuracy of the finding the same object in the next camera when the target object leaving from a camera.

Table 3. Performance comparison with full camera search in over camera tracking

Method/Dataset	Duke MTMC	NLPR_MCT_1	NLPR_MCT_2
full camera	40.9%	24.7%	28.1%
our method	87.6%	72.6%	84.2%

4 Conclusion

In this paper, a coarse-to-fine multi-camera network topology estimation method is proposed. We learn both the topological and temporal transition characteristics in the multi-camera network. Our approach does not require manual calibration and can automatically learn the transition relationship between the cameras. We test our method on several dataset. The experiments results demonstrate that our method can recovery the camera topology and the transition time between entries in multi-cameras surveillance video accurately than previous methods. And we also demonstrate that, in the over camera tracking application, our method can narrows the retrieval scope and plays a key role in finding out the object accurately and effectively.

Acknowledgments. This work is supported by the the Natural Science Foundation of China under Grant No.61572061, 61472020 and National 863 Program of China under Grant No.2015AA016403.

References

1. Makris, Dimitrios, Tim Ellis, and James Black. "Bridging the gaps between cameras." Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 2. IEEE, 2004.
2. T. Kinh, Gerald Dalley, and W. Eric L. Grimson. "Inference of non-overlapping camera network topology by measuring statistical dependence." Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Vol. 2. IEEE, 2005.
3. N. Chaowei, and Eric Grimson. "Recovering non-overlapping network topology using far-field vehicle tracking data." Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 4. IEEE, 2006.
4. M. Zehavit, Ilan Shimshoni, and Daniel Keren. "Multi-camera topology recovery from coherent motion." Distributed Smart Cameras, 2007. ICDSC'07. First ACM/IEEE International Conference on. IEEE, 2007.
5. X. Chen, Kaiqi Huang, and Tieniu Tan. "Learning the three factors of a non-overlapping multi-camera network topology." Pattern Recognition (2012): 104-112.
6. R. Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis & Machine Intelligence PP.99(2015):1-1.
7. H. John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.
8. Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking. E. Ristani, F. Solera, R. S. Zou, R. Cucchiara and C. Tomasi. ECCV 2016 Workshop on Benchmarking Multi-Target Tracking.
9. Xiao, Tong, et al. "Learning deep feature representations with domain guided dropout for person re-identification." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.
10. Tanaka, G., Suetake, N., Uchino, E.: Color Transfer Based on Normalized Cumulative Hue Histograms. JACIII. 14(2), 185-192 (2010)
11. Piti, F., Kokaram, A.: The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In: 4th European Conference on Visual Media Production, vol.1, pp. 1-9. IET (2007, November)
12. Xing, C., Ye, H., Yu, T., & Zhou, Z. (2016). Homogenous Color Transfer Using Texture Retrieval and Matching. Pacific Rim Conference on Multimedia(pp.159-168). Springer, Cham.