Multi-camera Tracking Exploiting Person Re-ID Technique

Yiming Liang and Yue Zhou(✉)

Institute of Image Processing and Pattern Recognition,
Shanghai Jiao Tong University, Shanghai, China
{liangyiming,zhouyue}@sjtu.edu.cn

Abstract. Multi-target multi-camera tracking is an important issue in image processing. It is meaningful to improve matching performance across cameras with high computational efficiency. In this paper, we apply high performance feature representation LOMO and metric learning XQDA in person re-identification across cameras to improve tracking performance. We also exploit direction information of trajectories to handle viewpoint variation. Experiments on DukeMTMCT dataset show that the proposed method improves tracking performance and is also competitive in running time.

Keywords: Multi-camera tracking · Multi-target tracking · Person re-identification

1 Introduction

Pedestrian tracking is a fundamental topic in computer vision. In the past decade, multi-target tracking has attracted lots of attentions, and large amounts of algorithms have been proposed to solve it. State-of the-art methods obtain impressive performance, but multi-target problem still needs further study, especially when it comes to a multi-camera problem.

Compared to single-target tracking, multi-target tracking mainly focuses on data association rather than appearance model [1]. Multi-target tracking methods are supposed to tackle difficult problems such as occlusions, detection failures, appearance similarity among targets and improving computational efficiency. Numerous approaches [1–3] have made contribution to improvement in performance and computational efficiency. [4] casts the problem of tracking multiple people as a graph partitioning problem and proposed Correlation Clustering by Binary Integer Programming (BIPCC). Since solving a BIP is NP hard, the size of the problem is reduced by a multi-stage cascade in which data is clustered according to space-time and appearance criteria. BIPCC obtains significant accuracy and high computational efficiency and is also employed in multi-camera tracking [5].

When it comes to tracking pedestrians across cameras, there are no strong space-time constraints if the camera pairs have no overlapping areas or we lack

D. Liu et al. (Eds.): ICONIP 2017, Part III, LNCS 10636, pp. 397–404, 2017.
https://doi.org/10.1007/978-3-319-70090-8_41

for real world information. Disregarding the weak space-time constraints, tracking across cameras can be consider as a person re-identification problem. To this end, we exploit effective approaches in person re-identification to improve the performance of multi-camera tracking.

Due to illumination changes, viewpoint variations, pose variations and occlusions, appearance-based person re-identification encounters difficulty in the past decade [6]. Most existing methods [7–11] concentrate their efforts on feature representation and metric learning. A feature representation named Local Maximal Occurrence (LOMO) together with a subspace and a metric learning method named Cross-view Quadratic Discriminant Analysis (XQDA) have been proposed in [12]. Experiments show that the combination of LOMO and XQDA obtains impressive performance as well as high computational efficiency. We find that applying LOMO and XQDA is useful in tracking multiple pedestrians across cameras.

In this paper, we propose a multi-target multi-camera tracking (MTMCT) approach named LXB (LOMO and XQDA based on BIPCC). We take tracking results within single camera as input of our system. The single-camera trajectories we used in experiments are accomplished by BIPCC. When it comes to tracking across cameras, LOMO is used to extract appearance features and XQDA is used to determine the similarity among targets. After we obtain the similarity matrix, another BIPCC is performed on targets across cameras. We also analyze the viewpoint, or direction, information of the targets. When comparing similarity between targets, bounding boxes indicating same or near viewpoints are given extra weightings. The proposed method is tested on a multi-target multi-camera tracking dataset named DukeMTMCT [5]. Experiments show that the proposed method improves the state-of-the-art scores on the dataset and is also competitive in running time.

2 Related Work

Numerous approaches [1–3] have been proposed to solve multi-target tracking problem. However, better performance usually requires higher computational complexity [4]. [13–15] approximate the solution to reduce computational complexity. [16,17] relax constraints in the BIP. [4] decomposes the problem into tracklets phase and trajectories phase and solves the sub-problems exactly with a BIP solver. This method has also been extended into multi-camera style in [5].

In most existing appearance-based methods [8,9,12] for person re-identification, different kinds of appearance features, mainly color and texture histograms, are combined in order to obtain higher robustness and matching performance. The feature histograms are weighted globally [18] or object-specifically [8] according to their capacity in distinguishing an object from a gallery. State-of-the-art approaches [1,7,12] deal with viewpoint changes by extracting better features or learning good metrics. The LOMO method in [12] maximizes the local occurrence of each SILTP and HSV histogram at the same horizontal location. A practical computation method for Cross-view QDA is also proposed.

[7] presents a descriptor that models a region as a set of multiple Gaussian distributions, and each Gaussian represents the appearance of a local patch. Then The characteristics of the Gaussian set are also described by a Guassian distribution. [1] trains a 16-layer VGGNet to extract appearance features.

Person re-identification methods mentioned above exhibit high performance on viewpoint variations. In contrast to them, we research utilizing appearance of a pedestrian from each available viewpoints under a camera to improve re-identification performance across cameras. It is reasonable to exploit information from as many viewpoints as possible. For instance, a pedestrian with frontal view under one camera can appear with back view under another camera [12]. If the pedestrian's clothes look different from the front and the back, matching across cameras becomes difficult and error-prone.

There have been numerous methods [19–22] for pedestrian direction estimation, most of which are aimed at the improvement of driver assistance systems. For example, [19] introduces a simple method to estimate walking direction of a pedestrian. Haar wavelets are used to generate feature vectors and SVMs with linear kernel are used to classify 16 directions. [20] is also a SVM-based method and estimates the discrete probability distribution of the directions. It also use a Hidden Markov Model to handle direction changes. For simplicity, our approach is also a SVM-based method and HOG is used to generate feature vectors.

3 MTMCT Framework

Section 3.1 introduces direction estimation of pedestrians. Section 3.2 describes how to exploit direction information in comparison between two trajectories. Section 3.3 introduces the BIP method we used in data association. Section 3.4 discusses feature representation and metric learning for person re-identification across cameras.

3.1 Direction Estimation of Pedestrian

We use HOG feature to generate feature vectors of pedestrians and train 8 SVM classifiers corresponding to 8 directions. The intervals between adjacent directions are 45°. Each of the classifiers generates an output at run time, and the direction corresponding to highest output is chosen for the pedestrian.

There is also a constraint that directions of a pedestrian can not change rapidly during tracking. As for a pedestrian trajectory, the difference between the viewpoints of adjacent frames is smoothed.

HOG are used to generate feature vectors for direction estimation. SVMs are used to classify the HOG feature vectors. According to our experiments results and [19], we chose a linear kernel function for better overall performance.

3.2 Exploiting Direction Information

For a pedestrian trajectory, the directions corresponding to each frame are estimated first. Then we respectively compute appearance features for the available

directions. For example, if a trajectory includes direction a, b and c, we extract 3 feature vectors respectively from the frames whose directions are a, b and c. Only appearance features with valid directions can be calculated when computing similarity between two trajectories.

When computing distance between features with directions of two trajectories, the directions are used to determine the weighting of the result. We denote the set of directions which appear in a trajectory as D. One D can contain at most 8 direction elements and at least 1 element. For trajectory A and B, their direction set are respectively D_A and D_B. Then the distance between A and B is defined as:

$$dir\_dist(A,B) = \frac{\sum_{i \in D_A,\ j \in D_B} w(i,j) \times dist(A_i, B_j)}{\sum_{i \in D_A,\ j \in D_B} w(i,j)} \tag{1}$$

where A_i is the feature vector of A corresponding to direction i, and B_j is the feature vector of B corresponding to direction j. $dist(\cdot,\cdot)$ can be any specific distance function between two feature vectors. $w(i,j)$ is the weighting for distance between two feature vectors with different directions. Since there are 8 directions in this system, the difference between two directions is at most 4. For direction i and j, the difference can be [0, 1, 2, 3, 4], and the corresponding $w(i,j)$ value are set as [1.0, 0.8, 0.6, 0.8, 1.0] according to performance in experiments. In Sect. 4.2 we illustrate the experiments and explain the choice among several groups of parameters.

3.3 Data Association

The system takes trajectories as input, and associate them into identities. We consider data association as a graph partitioning problem. Let V be a set of n trajectories. For $i, j \in V$, let $c_{ij} \in [-1, 1]$ represent the correlation between them. A higher correlation means they are more likely to belong to a same identity, and a lower value indicates that they are more unlikely to be the same person. Let the graph $G = (V, E, C)$ be a weighted graph on V. E is the set of edges connecting i and j in condition of c_{ij}. The system partitions V into sets in which trajectories belong to a same identity. The correlation clustering problem [23] is solved by a Binary Integer Program (BIP) on G:

$$\arg\max_{X} \sum_{(i,j) \in E} c_{ij} x_{ij} \tag{2}$$

subject to

$$x_{ij} \in \{0,1\} \quad \forall (i,j) \in E \tag{3}$$

If x_{ij} equals to 1, i and j are considered to be a same person, and vice versa. X is the set of all possible combinations of x_{ij}.

Solving this BIP is NP hard and the approximation is also hard [23,24]. Keeping the problems small can help improve efficiency. Similar to [4,13,25], we employ a sliding temporal window. The window moves forward by half of its temporal length. Solutions from previous overlapping windows are also fed into the

current window. Trajectories in a window are first divided into groups according to their appearance and, if available, space-time information. The division is to reduce the size of BIP and should be conservative to keep trajectories belonging to same person in the same group. Similar to [4], we employ simple k-means here. Then the solutions are computed in each group and in each temporal window.

3.4 Feature Representation and Metric Learning

Correlations in Sect. 3.3 are generated by appearance similarity and simple space-time criteria. As for appearance criteria, efficient and high-performance solutions in person re-identification can be employed here. LOMO and XQDA [12] show impressive performance as well as high computational efficiency in re-identifying person across cameras. LOMO maximizes the horizontal occurrence of local features and applies Retinex transform [26,27] to person images. XQDA learns a discriminant low dimensional subspace, and the metric is learned on the subspace. We estimate the direction information for each trajectory employing method in Sect. 3.1, and extract LOMO feature for each available direction. Then in each group and in each temporal window, features of all available trajectories are fed into a pre-trained XQDA model. Afterwards a distance matrix is generated. Elements in the matrix are transformed linearly into the range of $[-1, 1]$:

$$c_{ij} = -\frac{2}{d_{max} - d_{min}} \cdot \left(d_{ij} - \frac{d_{min} + d_{max}}{2} \right) \tag{4}$$

where c_{ij} and d_{ij} are the element in row i column j of correlation matrix C and distance matrix D respectively. d_{min} and d_{max} are the maximum and minimum value in D. Therefore, a longer distance is transformed into a lower correlation.

4 Experiments

We report the results on the DukeMTMCT [5] dataset. The DukeMTMCT dataset was captured from 8 synchronized cameras. It lasts for 85 min and contains more than 7,000 single camera trajectories and over 2,000 unique identities. Two camera pairs (2–8 and 3–5) have small overlapping areas, while the other cameras are disjoint. The running time was measured on a desktop PC with an Intel i7-6700 @ 3.40 GHz CPU.

4.1 Direction Estimation of Pedestrian

We randomly chose 1000 bounding boxes from ground truth training data of each camera in DukeMTMCT dataset [5], and got total 8000 samples. These bounding boxes are chosen from frame 49700 to 130000. All directions of the pedestrian samples are manual Annotated. 8 direction SVM classifiers are trained using 10-fold cross validation method. According to Table 1, This model is able to correctly classify most of the samples. In fact, most mis-classifications turn out to be adjacent directions of the correct ones. Therefore, the direction estimation model meets requirements.

Table 1. Mis-classification rates of direction estimation model

Direction (degree)	0	45	90	135	180	225	270	315
Mis-classification rate	0.088	0.063	0.057	0.097	0.087	0.066	0.051	0.089

4.2 Tracking Results on DukeMTMCT

Method in [5] is chosen to be the baseline for comparison, and we test our method from frame 130001 to 227540. The baseline method is also based on BIPCC and use striped color histograms together with simple temporal reasoning in matching targets across cameras. The input single camera trajectories are generated from [4]. All following experiments are based on these same input trajectories. The result is shown in Table 2. We also compare the performance of using different weighting for distance in Table 3. IDF1, IDP and IDR [5] are used to measure the performance.

Table 2. Performance comparison of our method

Tracker	IDF1 (%)	IDP (%)	IDR (%)	Time (seconds)
Baseline method	55.9	66.9	48.1	**2182.3**
Our method	**58.6**	**68.4**	**49.9**	2526.4
Baseline method+Direction	56.1	67.1	48.6	2362.9
Our method without direction	58.3	68.1	49.5	2345.5

As shown in Table 2, in comparison to the baseline method, our method yielded improved tracking score on the DukeMTMCT dataset. Both direction information and XQDA with LOMO help improve performance. However, improvement from direction information is not satisfactory considering the increased time consumption. In total, the increased time consumption of our method is acceptable.

Table 3. Performance comparison of different weighting for distance

Weighting	IDF1 (%)	IDP (%)	IDR (%)
Baseline method	55.9	66.9	48.1
$w = [1.0,\ 0.8,\ 0.6,\ 0.8,\ 1.0]$	**56.1**	**67.1**	**48.6**
$w = [1.0,\ 0.8,\ 0.6,\ 0.4,\ 0.2]$	56.0	67.0	48.4
$w = [1.0,\ 0.9,\ 0.8,\ 0.9,\ 1.0]$	56.0	67.0	48.3
$w = [1.0,\ 0.5,\ 0.01,\ 0.01,\ 0.01]$	55.3	66.1	47.5

As shown in Table 3, when the difference between two directions is as large as 3 or 4, higher weighting yield higher performance. It may be because the

appearance of a person are usually more similar between opposite viewpoints than orthogonal viewpoints. Therefore, $w = [1.0,\ 0.8,\ 0.6,\ 0.8,\ 1.0]$ is finally chosen in our system.

5 Conclusion

In this paper, we innovatively import the LOMO and XQDA method and direction information to match person across cameras in MTMCT. Experiments prove that our approach outperforms the original baseline and is also competitive in running time.

Acknowledgments. This work is supported by National High-Tech R&D Program (863 Program) under Grant 2015AA016402.

References

1. Sadeghian, A., Alahi, A., Savarese, S.: Tracking the untrackable: learning to track multiple cues with long-term dependencies. arXiv preprint arXiv:1701.01909 (2017)
2. Yu, S.I., Meng, D., Zuo, W., Hauptmann, A.: The solution path algorithm for identity-aware multi-object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3871–3879 (2016)
3. Yoon, J.H., Lee, C.R., Yang, M.H., Yoon, K.J.: Online multi-object tracking via structural constraint event aggregation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1392–1400 (2016)
4. Ristani, E., Tomasi, C.: Tracking multiple people online and in real time. In: Asian Conference on Computer Vision, pp. 444–459 (2014)
5. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, pp. 17–35 (2016)
6. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based person reidentification in camera networks: problem overview and current approaches. J. Ambient Intell. Hum. Comput. **2**(2), 127–151 (2011)
7. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical Gaussian descriptor for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1363–1372 (2016)
8. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: what features are important? In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7583, pp. 391–401. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33863-2_39
9. Farenzena, M., Bazzani, L., Perina, A., Cristani, M., Murino, V.: Person reidentification by symmetry-driven accumulation of local features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2367 (2010)
10. Li, W., Zhao, R., Xiao, T., Wang, X.: DeepReID: deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 152–159 (2014)

11. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by discriminative selection in video ranking. IEEE Trans. Pattern Anal. Mach. Intell. **38**(12), 2501–2514 (2016)
12. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
13. Shafique, K., Shah, M.: A noniterative greedy algorithm for multiframe point correspondence. IEEE Trans. Pattern Anal. Mach. Intell. **27**(1), 51–65 (2005)
14. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: Computer Vision and Pattern Recognition, pp. 1273–1280 (2011)
15. Roshan Zamir, A., Dehghan, A., Shah, M.: GMCP-tracker: global multi-object tracking using generalized minimum clique graphs. In: Computer Vision, CECCV, pp. 343–356 (2012)
16. Jiang, H., Fels, S., Little, J.J.: A linear programming approach for multiple object tracking. In: Computer Vision and Pattern Recognition, pp. 1–8 (2007)
17. Butt, A.A., Collins, R.T.: Multi-target tracking by Lagrangian relaxation to min-cost network flow. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1853 (2013)
18. Mignon, A., Jurie, F.: PCCA: a new approach for distance learning from sparse pairwise constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2666–2672 (2012)
19. Shimizu, H., Poggio, T.: Direction estimation of pedestrian from multiple still images. In: Intelligent Vehicles Symposium, pp. 596–600. IEEE (2004)
20. Gandhi, T., Trivedi, M.M.: Image based estimation of pedestrian orientation for improving path prediction. In: Intelligent Vehicles Symposium, pp. 506–511. IEEE (2008)
21. Flohr, F., Dumitru-Guzu, M., Kooij, J.F., Gavrila, D.M.: Joint probabilistic pedestrian head and body orientation estimation. In: Intelligent Vehicles Symposium Proceedings, pp. 617–622. IEEE (2014)
22. Tao, J., Klette, R.: Part-based RDF for direction classification of pedestrians, and a benchmark. In: Asian Conference on Computer Vision, pp. 418–432 (2014)
23. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. In: Foundations of Computer, Science, pp. 238–247 (2002)
24. Tan, J.: A note on the inapproximability of correlation clustering. Inf. Process. Lett. **108**(5), 331–335 (2008)
25. Javed, O., Shafique, K., Rasheed, Z., Shah, M.: Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Comput. Vis. Image Underst. **109**(2), 146–162 (2008)
26. Jobson, D.J., Rahman, Z.U., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE Trans. Image Process. **6**(3), 451–462 (1997)
27. Jobson, D.J., Rahman, Z.U., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans. Image Process. **6**(7), 965–976 (1997)