A NOVEL SOLUTION FOR MULTI-CAMERA OBJECT TRACKING

Weihua Chen *Lijun Cao* *Xiaotang Chen* *Kaiqi Huang*

Center for Research on Intelligent Perception and Computing,
National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences

ABSTRACT

The traditional multi-camera object tracking contains two steps: single camera object tracking (SCT) and inter-camera object tracking (ICT). The ICT performance strongly relies on the great results of SCT. In practice, most of current SCT methods are unperfect and products much more fragments. In this paper, a novel solution using a global tracklet association is proposed, which can provide a good ICT performance when the SCT results are not perfect. The proposed solution is also available in non-overlapping views through a new tracklet representation and experiments shows the effectiveness of the proposed novel solution in real scene.

Index Terms— Multi-camera object tracking, global tracklet association, PMCSHR

1. INTRODUCTION

Nowadays thousands of cameras in the world are collecting a huge amount of data on a daily basis for surveillance. It becomes increasingly important to develop related methods that process these data automatically. Multi-camera object tracking plays an important role in video surveillance. The goal is not only to monitor pedestrians, but also to extract useful information for other tasks at semantically higher levels, such as personal behavior analysis.

Today most researchers split multi-camera object tracking into two tasks, shown in Fig. 1 (Solution A): single camera object tracking (SCT) [1, 2, 3, 4] and inter-camera object tracking (ICT) [5, 6, 7, 8, 9, 10, 11]. SCT is used to get multi-object trajectories in a single camera view, while ICT aims to connect these trajectories across multiple camera views. The input of ICT is the output of SCT, which causes ICT strongly depends on the performance of SCT methods. However, current SCT methods [1, 2, 3, 4] are not as perfect as they wish. There're too many fragments in the results of SCT task, which fails to meet the requirements as the inputs for ICT. These fragments may bring new problems into ICT task, such as wrong matching problem, which means two targets in Camera 2 are matched to different tracklets of a same target in Camera 1 (see Fig. 2 (a)), and tracklet missing problem, as that some tracklets of a target are missing during inter-camera tracking (see Fig. 2 (b)). In this paper, the proposed method integrates SCT and ICT together to avoid these problems.

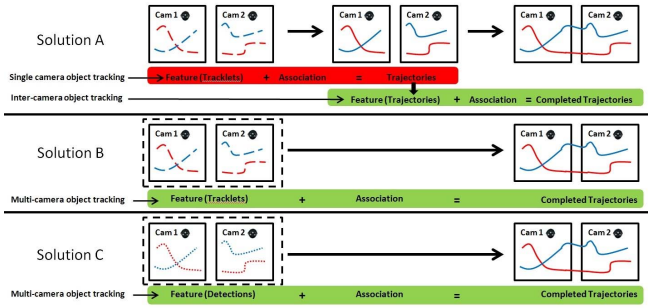
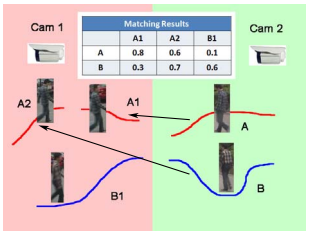


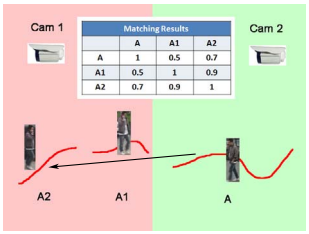
Fig. 1. Three kinds of solutions for multi-camera tracking

Multi-camera object tracking in the proposed method is considered as a global tracklet association under a panoramic view (see Fig. 1 (Solution B)). And different tracklets in the global tracklet association are treated differently according to the cameras they belong to. This framework provides a new solution to multi-camera object tracking when the SCT performance is not good enough for further ICT process. Frankly, its local performance in a specific camera view may be as fragmentary as that of the traditional SCT methods, but it overcomes the new problems emerging in ICT when SCT is not good and offers a better ICT performance as a compensation. In practice, a better ICT has strong practical significance, for a video surveillance system, it's more important to locate the objects in wide area than single scene.

Some researches [12, 13, 14] also pay their attentions on this framework. They mainly follow a tracking-by-detection paradigm and form the association graph (see Fig. 1 (Solution C)). Yu [12] proposes a nonnegative discretization solution for data association and identifies people across different cameras by face recognition. While for real scene with objects in a distant view, faces are too small to be recognized. Hofmann [13] uses a global min-cost flow graph and connects the different-view detections through their overlapping locations in a world coordinate space, which can't solve the non-overlapping view problem. In this paper, the proposed novel solution uses fragmentary tracklets as the inputs instead of ob-



(a) Wrong matching



(b) Tracklet missing

Fig. 2. Two matching problems. Blue and red lines indicates two targets and arrows show the best matching. Target B is matched to tracklet A2 wrongly in (a). Tracklet A1 is missing in (b).

ject detections. A piecewise major color spectrum histogram representation (PMCSHR) taking advantage of the periodicity of pedestrian walking is introduced to achieve tracklet matching across multiple camera views, which makes the proposed solution suitable for the tracking in non-overlapping views.

Our contributions to multi-camera object tracking are in three aspects: 1) a novel solution for multi-camera tracking in non-overlapping views is proposed; 2) a global tracklet association is introduced to solve the unprefect input from SCT for ICT; 3) a PMCSHR is used to represent tracklets and experiments show its effectiveness and discriminative power.

The rest of the paper is organized as follows: The global tracklet association graph is described in Section 2. The PMCSHR is introduced in Section 3. The experimental results in Section 4 shows the effectiveness of the proposed solution.

2. MAP ASSOCIATION TRACKING

In this paper, the global tracklet association is modeled as a global maximum a posteriori (MAP) problem, which can be mapped into a cost-flow network and solved by a min-cost flow algorithm.

2.1. Tracklets

Pedestrian are tracked via the MAP association by first extracting a set of tracklets L. A head-shoulder detector [15] and an AIF tracker [16] are first used to get all the tracklets from each camera. During the target tracking by the AIF, a confidence α_t is got to evaluate the accuracy of a tracking result [16] in frame t. If the confidence is low enough $\alpha_t < 0.2$, the tracker is considered to be lost. Then all confidence values of the target are recorded during tracking and the average value c is computed as the likelihood of tracklet i.

$$c_i = \frac{\Sigma_{j=t_s}^{t_e} \alpha_j}{(t_e - t_s)} \tag{1}$$

where t_s and t_e are the start and end frame for tracklet i.

So all the tracklets from all cameras are obtained, $L = \{l_1, l_2, ..., l_N\}$ and N is the total number of tracklets in all cameras, where each tracklet $l_i = [x_i, c_i, s_i, t_i, a_i]$ consists of position, likelihood, camera view, time stamp and appearance information respectively.

2.2. MAP Formulation

The proposed solution shared the same MAP formulation with Zhang [3]. The difference is that the input in the proposed solution are tracklets instead of object detections. And the association aims to solve the wrong matching and tracklet missing problems in ICT, while Zhang [3] applies it on SCT. In our approach, a single trajectory hypothesis is defined as an ordered list of target tracklets, i.e. $T_i = \{l_{i_1}, l_{i_2}, ..., l_{i_k}\}$ where $l_{i_k} \in L$. The association trajectory hypothesis T is defined as a set of single trajectory hypothesis, i.e. $T = \{T_i\}$. The objective of data association is to maximize the posteriori probability of T given the tracklets set L under the non-overlap constraints [3]:

$$T^* = \arg\max_T \prod_i P(l_i|T) \prod_{T_k \in T} P(T_k) \tag{2}$$

$$T_i \cap T_j = \emptyset, \forall i \neq j \tag{3}$$

$P(l_i|T)$ is the likelihood of tracklet l_i, and set to c_i. The prior $P(T_k)$ is modeled as a Markov chain containing transition probabilities $\prod P(l_{k_{i+1}}|l_{k_i})$ of all tracklets in T_k [13].

2.3. Min-cost Flow Solution

The MAP association can be solved by a min-cost flow [17]. In its cost function, the proposed solution distinguishes different tracklets and increases the weights of those from different cameras. In the min-cost flow graph, a vertex i is defined for each tracklet l_i with a likelihood c_i and edge weights reflect the prior $P(T_k)$. The cost e_{ij} per flow from i to j indicates the negative affinity, which we decompose into probabilities in continuity of motion, time and appearance.

$$e_{ij} = -\log(P_m \cdot P_t \cdot P_a) \tag{4}$$

The probabilities P_m and P_t are obtained using the same methods from [4] and [17] respectively, and the appearance probability P_a is the key part in the proposed solution to distinguish different tracklets,

$$P_a = \begin{cases} Dis(A_i, A_j) & \text{if } s_i = s_j \\ \lambda Dis(A_i, A_j) & \text{if } s_i \neq s_j \end{cases} \tag{5}$$

where A_i is the PMCSHR for tracklet l_i (seen in Section 3.1), λ is a compensation factor for the similarity of tracklets across camera views, and $Dis(x, y)$ is the similarity measurement for two tracklets' PMCSHRs.

From (5), a difference is made between tracklets from the same camera and different cameras. In other word, we emphasize the importance of tracklets across camera views.

3. PMCSHR AND ITS MEASUREMENT

In this section, the PMCSHR is introduced first. Then a new measurement based on minimum uncertainty gap (MUG) is proposed to compute the matching result of two tracklets' PMCSHRs.

3.1. PMCSHR

The PMCSHR is calculated based on the MCSHR [5] which obtains the major colors of a target based on an online k-means clustering algorithm. When computing the MCSHR for a tracklet, the normal way is to integrate the histograms from all the frames together,

$$H_i = \sum_{j=1}^{m_i} h_j \tag{6}$$

where h_j is the MCSHR for tracklet i in the jth frame and H_i is the incremental MCSHR [5] for the whole tracklet i. m_i is the length of tracklet i.

Pedestrian as a non-rigid target is not much like vehicle in tracking, and its MCSHR is still not easy for tracking and identification. However the pedestrian has its own special characteristic. When a person walks, his walking has obvious periodic characteristic that arms and legs wave around periodically in a certain time. It assumes that people always walk at a constant speed. The proposed solution is to find this periodic time p_i and use it to segment tracklets.

As a preparation, all MCSHRs $\{h_1, h_2, ..., h_{m_i}\}$ for tracklet i are obtained and the similarity $\Lambda_{k,j}$ between any pair h_k and h_j is computed [5]. The idea is to compute all the possible periodic times and find the best one. For a certain periodic time $p_i = t$, the similarity $\Lambda_{j,j+t}$ between h_j and its next periodic h_{j+t} is collected for every frame j in tracklet, and the average similarity is considered as the value which determines the validity of this periodic time t.

$$p_i = \arg\max_t \frac{1}{m_i - t} \sum_{j=1}^{m_i - t} \Lambda_{j,j+t} \quad \forall t \in [\gamma, m_i) \tag{7}$$

The set $[\gamma, m_i/2)$ is used to estimate the possible range of t, and γ is set to 15. If γ is too small, the nearby frames have a strong similarity which causes (7) to a false maximum. After calculation, p_i is the best periodic time for tracklet i. Then the tracklet i can be evenly segmented into pieces with the length p_i (except the end part). For each piece, the incremental MCSHR is computed [5]. The PMCSHR for tracklet i is represent as:

$$A_i = \{H_1, H_2, ..., H_{d_i}\} \tag{8}$$

where $d_i = \lceil \frac{m_i}{p_i} \rceil$ is the number of pieces the tracklet i segmented.

3.2. The MUG Measurement

During matching, the uncertainty of the likelihood between tracklets is as important as the likelihood itself. We try to find the relationship which minimizes uncertainty of the likelihood between two tracklets besides the traditional two directional similarity (TDS) measurement [5] using color distances in RGB space. As a result, the goal of tracklets matching problem is defined as to find the best state that maximizes the average bound of the similarity and minimizes the gap between bounds of the similarity, sharing the same idea with [18]. As obtained in 3.1, each piece is considered as a complete and independent period to represent their tracklet. Then the lowest and highest similarity between two pieces from two different tracklets are found and defined as the lower and upper bounds of the similarity for these two tracklets.

$$Dis(A_i, A_j) = 1 - \frac{\max Sim(H_u, H_v) - \min Sim(H_f, H_g)}{\max Sim(H_u, H_v) + \min Sim(H_f, H_g)} \tag{9}$$

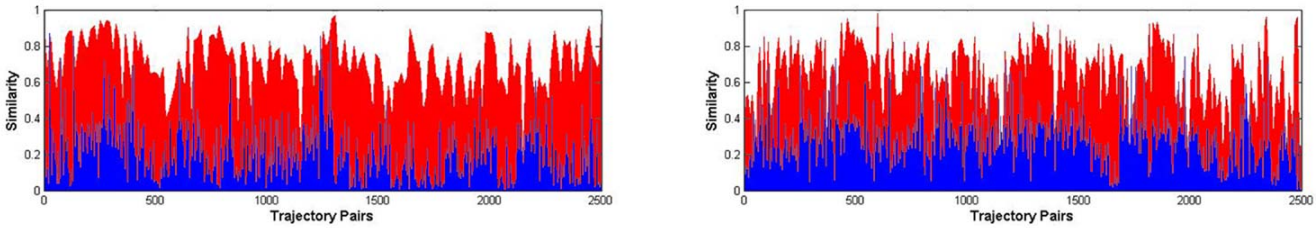
where $u, f \in d_i$ and $v, g \in d_j$. This distance not only considers the average likelihood, but also takes the uncertainty of the likelihood into account.

4. EXPERIMENTS

The experiments are tested on NLPR_MCT dataset [19] that consists of four sub-datasets. Every sub-dataset includes 3-5 cameras and has a different situation according to the number of people, ranging from 14 to 255, and the level of illumination changes and occlusions. All the videos are nearly 20 minutes (except Dataset 3) and recorded under non-overlapping views in real scenes during daily time, which make the dataset a good representation of different situations in normal life. In this paper, we test the proposed method on Dataset 4, because of its complex environment and its long trajectories. It contains five cameras and the length of each video is 25 minutes with a rate of 25 fps. First, the effectiveness of the PMCSHR and the MUG measurement are first evaluated. Then the tracking results using the proposed solution is presented. It is shown that the proposed solution achieves ICT well when the SCT process causes much more fragments than ground truth.

4.1. PMCSHR and MUG

As five cameras used, there are 10 (C_5^2) two-camera pairs. The experiments are done on all 10 pairs and sum them up as final results. Here two matching similarities are defined for evaluation: the positive similarity (PS) and the negative similarity (NS). The PS is the average similarity between every trajectory pair from two different cameras which has a link in the ground truth, and the NS means that between non-linked pairs. The higher PS with a lower NS indicates more discriminative power during matching. The input in this experiment



(a) PMCSHR+MUG

(b) MCSHR+TDS [5]

Fig. 3. The PS and NS Results of the PMCSHR+MUG and the MCSHR+TDS [5]. The red and blue bars indicate the PSes of 2483 positive pairs and the NSes for 2483 negative pairs. For a clear comparison, we only shows 2483 negative pairs which are selected randomly from 41805 negative pairs.

is the ground truth instead of our tracking results.

The average PS and NS of the PMCSHR+MUG are 0.7 and 0.1248, while the average PS and NS of MCSHR+TDS [5] are 0.6247 and 0.1425. That indicates the proposed PMCSHR using MUG measurement has a more discriminative power to measure the similarity between any two trajectories. From the results shown in Fig. 3, it can be seen that it's more easier finding a threshold to distinguish PS and NS in the proposed PMCSHR. In this experiment, the obtained average period of pedestrian walking is 24 frames, which implies people finish a cycle of walking in nearly 1 second. It matches the pedestrian walking period in common sense.

4.2. Our Approach

This experiment aims to evaluate the effectiveness of the proposed solution on multi-camera tracking. The advantage of the proposed solution is to improve the ICT performance under an unperfect SCT result. The proposed solution is compared with the MCSHR+TDS [5] and the PMCSHR+MUG without a MAP association process. There's only a Hungary algorithm in them as a simple ICT association. A quantitative evaluation extended from [11] is used to evaluate the performance, which consists of four parts: crossing fragments (X-Frag), crossing ID switches (X-IDS), success rate (SR) and failure rate (FR). Crossing fragments and crossing ID switches are referred to [11]. The success rate is a ratio of the positive linked pairs among all the linked pairs, while the failure rate means that of the negative linked pairs. As an input, the pedestrian detector and single object tracking are the methods mentioned in Section 2.1, which are comparable to their corresponding state-of-the-arts, but still product much more fragments. After single object tracking, 543 tracklets are obtained, while there are only 333 tracklets in ground truth. The compensation factor λ is set to 2.5 by experience.

From Table 1, PMCSHR+MUG has a lower X-Frag and a lower X-IDS because of its more discriminative power. And a sharp fall of our method (PMCSHR+MUG+MAP) in X-Frag can be explained by that many tracklets left by Hungary algorithm are re-linked in the MAP association. It implies the proposed method partly solves the tracklet missing problem. In another aspect, the less X-IDS and the higher SR can be seen as an indication that the wrong matching problem is overcame to some extent. As a result, the proposed solution can be considered more reliable for ICT under a unperfect SCT.

	X-Frag	X-IDS	SR	FR
input tracklets	543	0	0	0
[5]	86	137	0.793	0.207
PMCSHR+MUG	93	109	0.831	0.169
PMCSHR+MUG+MAP	19	84	0.869	0.131

Table 1. Tracking results using different approaches under an unperfect SCT which causes much more fragments. It shows that the proposed solution achieve the best performance.

5. CONCLUSION

A novel solution for multi-camera tracking in non-overlapping views is proposed. It partly solves the wrong matching and tracklet missing problems emerging in ICT through a global tracklet association when the SCT performance is not good enough. A novel PMCSHR with a MUG measurement is introduced to represent tracklets and the experiments shows its effectiveness and discriminative power.

6. ACKNOWLEDGEMENT

This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302), National Natural Science Foundation of China (Grant No. 61322209 and Grant No. 61175007), the National Key Technology R&D Program (Grant No. 2012BAH07B01).

7. REFERENCES

[1] C.H. Kuo and R. Nevatia, "How does person identity recognition help multi-person tracking?," in *CVPR*. IEEE, 2011, pp. 1217–1224.

[2] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *CVPR*. IEEE, 2009, pp. 2953–2960.

[3] L. Zhang, Y. Li, and R. Nevatia, "Global data association for multi-object tracking using network flows," in *CVPR*. IEEE, 2008, pp. 1–8.

[4] B. Yang and R. Nevatia, "An online learned crf model for multi-target tracking," in *CVPR*. IEEE, 2012, pp. 2034–2041.

[5] M. Piccardi and E.D. Cheng, "Multi-frame moving object track matching based on an incremental major color spectrum histogram matching algorithm," in *CVPR*. IEEE, 2005, p. 19.

[6] X. Chen, K. Huang, and T. Tan, "Object tracking across non-overlapping views by learning inter-camera transfer models," *Pattern Recognition*, vol. 47, pp. 1126–1137, March 2014.

[7] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," in *ICCV*. IEEE, 2003, vol. 2, pp. 952–957.

[8] R. Hamid, R.K. Kumar, M. Grundmann, K. Kim, I. Essa, and J. Hodgins, "Player localization using multiple static cameras for sports visualization," in *CVPR*. IEEE, 2010, pp. 731–738.

[9] Z. Wu, N.I. Hristov, T.L. Hedrick, T.H. Kunz, and M. Betke, "Tracking a large number of objects from multiple views," in *ICCV*. IEEE, 2009, pp. 1546–1553.

[10] H. Jiang, S. Fels, and J.J. Little, "A linear programming approach for multiple object tracking," in *CVPR*. IEEE, 2007, pp. 1–8.

[11] C.H. Kuo, C. Huang, and R. Nevatia, "Inter-camera association of multi-target tracks by on-line learned appearance affinity models," *ECCV*, vol. 6311, pp. 383–396, 2010.

[12] S.I. Yu, Y. Yang, and A. Hauptmann, "Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization," in *CVPR*. IEEE, 2013.

[13] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," in *CVPR*. IEEE, 2013.

[14] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn, "Branch-and-price global optimization for multi-view multi-target tracking," in *CVPR*. IEEE, 2012, pp. 1987–1994.

[15] M. Li, Z. Zhang, K. Huang, and T. Tan, "Rapid and robust human detection and tracking based on omega-shape features," in *ICIP*. IEEE, 2009, pp. 2545–2548.

[16] W. Chen, L. Cao, J. Zhang, and K. Huang, "An adaptive combination of multiple features for robust tracking in real scene," in *ICCVW*. IEEE, 2013, pp. 129–136.

[17] J. Liu, P. Carr, R.T. Collins, and Y. Liu, "Tracking sports players with context-conditioned motion models," in *CVPR*. IEEE, 2013.

[18] J. Kwon and K.M. Lee, "Minimum uncertainty gap for robust visual tracking," in *CVPR*. IEEE, 2013, pp. 2355–2362.

[19] "Nlpr_mct dataset," http://mct.idealtest.org/Datasets.html.