An equalised global graphical model-based approach for multi-camera object tracking

Lijun Cao, Weihua Chen, Xiaotang Chen, Member, IEEE, Shuai Zheng, and Kaiqi Huang, Senior Member, IEEE

Abstract—Multi-camera non-overlapping visual object tracking system typically consists of two tasks: single camera object tracking and inter-camera object tracking. Since the state-of-the-art approaches are yet not perform perfectly in real scenes, the errors in single camera object tracking module would propagate into the module of inter-camera object tracking, resulting much lower overall performance. In order to address this problem, we develop an approach that jointly optimise the single camera object tracking and inter-camera object tracking in an equalised global graphical model. Such an approach has the advantage of guaranteeing a good overall tracking performance even when there are limited amount of false tracking in single camera object tracking. Besides, the similarity metrics used in our approach improve the compatibility of the metrics used in the two different tasks. Results show that our approach achieve the state-of-the-art results in multi-camera non-overlapping tracking datasets.

Index Terms—Multiple non-overlapping camera object tracking, equalized global graphical model, multiple camera object tracking

I. INTRODUCTION

TRACKING objects of interest is an important and challenging problem in intelligent visual surveillance systems [1]. visual object tracking [2] is a long-standing problem in computer vision, there are a great amount of efforts made in visual object tracking within single cameras [3], [4]. Since the visual surveillance systems provide huge amount of video streams, it is desirable to develop algorithms to track those objects of interest instead of human. In intelligent visual surveillance systems [5], [6], due to the finite camera field of view, it is difficult to observe the complete trajectory of objects of interest in wide areas with only one camera. Hence, it is desired to enable the intelligent visual surveillance system to track the objects of interest within multiple cameras [7]. In addition, for practical considerations in installing an intelligent visual surveillance system, it is usually to have the cameras with no overlapping areas. Thus, the intelligent visual surveillance system should be able to track objects of interest across multiple non-overlapping cameras. In this paper, we focus on addressing the problem of tracking objects of interest across multiple non-overlapping cameras.

As shown in Fig. 1, previous approaches address the challenges in this multi-camera object tracking (MCT) problem in two different directions: single camera object tracking (SCT) [8], [9], [10] and inter-camera object tracking (ICT) [11], [12], [13]. In the direction of SCT, approaches [8], [9], [10] attempted to compute the trajectories of multiple objects from a single camera view. Alone this line, approaches [11], [12], [13] aim to find the correspondences among those trajectories across multiple camera views. Those ICT approaches often use the trajectories results from single camera object tracking to achieve the data association, hence the overall tracking system is brittle and overall performance depends on the performance of the single camera object tracking module. The direct disturbance of false positives and the fragments also bring problems into ICT module, such as wrong matching problem. i.e. 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, i.e. some tracklets of a target are missing during inter-camera tracking (see Fig. 2 (b)). We address these problems by considering to integrate the two separate modules and jointly optimise them.

In order to achieve this, we develop a global multi-camera object tracking approach [14] it integrates two tasks together via an equalised global graphical model to avoid these "inevitable" problems and aims to improve the overall performance of multi-camera object tracking.

This proposed approach is developed respecting to the two criterions of the two tasks:

- Single camera object tracking criterion: obtaining more completed pedestrian trajectories in single camera which can rebuild their exact historical paths in each scene.
- Inter-camera object tracking criterion: achieving the inter-camera matching of pedestrian trajectories that can help to locate pedestrian in a wide area.

As shown in Fig. 1 (Solution A), SCT and ICT share a similar data association framework: a graph modeling with an optimisation solution. The data association inputs in the single camera object tracking module are the initial observations, such as detections or tracklets, and the outputs are the integrated trajectories in each single camera (known as mid-term trajectories). These mid-term trajectories are then used as inputs to achieve the data association in inter-camera object tracking, and the outputs of the ICT approaches are the final integrated trajectories in multi-cameras (known as final trajectories). To integrate these two data association together,
the straightforward idea is to establish a new data association which takes initial observations as inputs and outputs the final trajectories directly. Nevertheless, a new problem arises, i.e. how to measure the similarities between two observations in the new graph. Some of the trajectories are from the observations in the same camera, and others are from that in different cameras. Hence, under the same similarity metric, the average similarity score between two observations in different cameras are often lower\(^2\) than that in the same camera, because the appearance and the spatio-temporal information of objects are less reliable in ICT than those in SCT due to many factors (camera settings, viewpoints and lighting conditions). In this case, it encourages the graph gives priority to linking the observations following the edges in the same camera instead of those across cameras. These bring new problems, i.e. how to distinguish the similarities in a same camera from those in different cameras, and how to balance them in one graph?

In this paper, we improve the similarity metrics, where we treat the similarities of SCT and ICT differently and equalizes them in a global graph. A minimum uncertain gap (MUG) \(^1\) is adopted to establish the improved similarity metrics. Thanks to the MUG, the similarity scores in both SCT and ICT are equalised under the proposed global graphic model.

The contributions of this paper are as follows.

1) a global graphic model for multi-camera object tracking is presented which integrates SCT and ICT tasks together to avoid the “inevitable” problems;
2) the similarity metrics are improved to equalise the different similarities in two tasks and unify them under one graph;
3) the proposed approach is experimented on a comprehensive evaluation criteria which clearly shows its effectiveness compared with the traditional two-step MCT framework.

II. RELATED WORK

Using a graphic model is an efficient and effective way to solve the data association problem in the multi-camera visual object tracking. It follows two parts, a graph modeling and an optimisation solution. Graph modeling is to form a solvable graphic model by input observations (detections, tracklets, trajectories or pairs). It includes nodes, edges and weights.

The optimisation solution is used to solve the graph to achieve optimal or suboptimal solutions. The difference is that single camera object tracking (SCT) emphasizes particularly on the graph and the optimisation solution, i.e. how to build a more efficient or more discriminative graph. While inter-camera object tracking (ICT) focuses on nodes, edges and weights, it prefers getting a more effective feature representation, especially for object re-identification (Re-ID). The ICT has a more complex and more sophisticated representation or similarity metrics (i.e. a transition matrix), but with a simpler graphic model. The proposed approach takes advantages of both SCT and ICT. The proposed similarity metrics are extended from a classical Re-ID method \(^13\), \(^20\) and the global graph model takes advantage of a state-of-the-art SCT approach \(^21\).

This section reviews the related approaches for each part of SCT, ICT and MCT. Section 2.1 reviews the single camera multi-object tracking. Section 2.2 discusses the inter-camera object tracking with a brief introduce of object re-identification. Section 2.3 reviews some other multi-camera object tracking approaches that take both SCT and ICT into account.

A. Single Camera Object Tracking

In single camera multi-objects tracking (SCT), the prediction of the spatio-temporal information of objects is more reliable and the appearance of objects does not have much variation during tracking. This makes the SCT task less challenge than the ICT task. i.e. For some less challenge videos, a simple appearance representation (e.g. color histogram \(^22\), \(^23\), \(^24\)) works well. The graphic model are often used to solve different problems, such as occlusion \(^25\), \(^26\), crowd \(^24\), \(^27\) and interference of appearance similarity \(^28\), \(^29\). However, for challenge videos, these approach led to frequent id-switch errors and trajectory fragmentation.

Existing approaches in SCT usually follows a data association-based tracking framework, which links short tracklets \(^18\), \(^23\), \(^30\) or detection responses \(^31\), \(^32\), \(^33\) into trajectories by global optimization based on kinds of features, such as motion (position, velocity) and appearance (color, shape). The improvements are always along two aspects: the graph model and the optimization solution. Some researchers focus on developing a new graph model for their tracklets or detections and aim to solve a specific problem. In Possegger et al. \(^26\), a geodesics method is adopted for data association to handle the occlusion problem. Dicle et al. \(^28\) uses motion dynamics to solve generalized linear assignment when the appearance similarity exists. Other works in SCT focus on the improvement of the optimization solution framework, such as continuous energy minimization \(^34\), linear programming \(^35\), CRF \(^36\) and the mixed integer program \(^37\). Zhang et al. \(^21\) proposes a maximum a posteriori (MAP) model to solve the data association of the multi-object tracking efficiency, while Yang et al. \(^36\) utilizes an online CRF approach to handle the optimization with the benefit of distinguishing spatially close targets in similar appearances. These approaches can partly yield id-switches and trajectory fragmentation, but separated optimisation makes
them suffering from leaving many fragmentation and false positives for ICT step.

B. Inter-camera Object Tracking (ICT)

Inter-camera tracking is more challenge than SCT because of its greater dramatic changes in appearance caused by many factors (camera settings, viewpoints and lighting conditions) and less reliable spatio-temporal information in different camera views. As a result, how to learn a discriminative and invariant feature representation and suitable similarity metrics are the mainly problems in ICT.

Most ICT works solve this problem from multi-camera calibration [38], [39], [40] and feature cues [41], [42], [43], [44], [45]. For multi-camera calibration, as an immobile information, the approaches in this aspect always project the multiple scenes into a 3D coordinate system, and achieve the matching using projected position information. Hu et al. [39] adopts a principal axis-based correspondence to achieve the calibration. For feature cues, most approaches deal the matching with improved appearance or spatio-temporal information. Kuo et al. [42] applies MIL on learning appearance affinity model, while Matei et al. [43] integrates appearance and spatio-temporal likelihoods within a multi-hypothesis framework.

Considering the graph model-based approaches, a K-camera ICT data association can be treated as a K-partite graph matching problem. It is difficult to get the optimal solution, but there’re many approaches to get the suboptimal solutions, e.g. the weighted bipartite graph [11], the Hungarian algorithm [46] and the binary integer program [47]. The K-partite idea holds an assumption that each camera has had a perfect tracklets 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, even the across-camera information may provide some useful feedbacks for each specific camera. But it overcomes the new problems as the ICT approaches have a common assumption that the single camera object tracking results are perfectly done and the trajectories in single cameras are all true positive and integrated well, which is hard to be achieved today.

C. Multi-camera Object Tracking (MCT)

Let’s back to MCT, a good MCT is the ultimate goal for any researcher in tracking. Most MCT methods follow the two-step framework, a SCT algorithm plus a ICT algorithm. In the Multi-Camera Object Tracking Challenge [53] in ECCV 2014 visual surveillance and re-identification workshop, methods of most participating teams are two-step approach. The winner USC-Vision team uses a state-of-the-art SCT method [32] and an improved ICT method [41], which is comparable with the state-of-the-art method [42].

Besides two-step approaches, there’re some multi-camera object tracking approaches [54], [55], [56], [57] concentrating on integrating the process of SCT and ICT into one global graph as this paper does. They mainly follow a tracking-by-detection paradigm and form a global association graph (see Fig. 1 (Solution C)). Yu et al. [55] 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 et al. [57] 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 camera problem.

In this paper, the proposed solution uses fragmentary tracklets as the inputs instead of object detections. It considers multi-camera object tracking as a global tracklet association under a panoramic view (see Fig. 1 (Solution B)). And the similarities of different tracklets in the global tracklet association are treated differently according to the cameras the tracklets 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, even the across-camera information may provide some useful feedbacks for each specific camera. 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, its more important to locate the objects in wide area than single scene.

III. EQUALISED GLOBAL GRAPHIC MODEL

The proposed approach aims to unify SCT and ICT into one global data association. The inputs are the observations and the
outputs are final trajectories. The data association is modeled as a global maximum a posteriori (MAP) problem which shares the same MAP formulation with Zhang et al. [21]. The difference is that the input in the proposed solution are tracks instead of object detections. And the association aims to solve the wrong matching and tracklet missing problems in ICT, while Zhang et al. [21] applies it on SCT. We outline the variable definitions in Tab. I.

In our approach, a single trajectory hypothesis is defined as an ordered list of target tracklets, i.e. $\Gamma_i = \{l_{i1}, l_{i2}, ..., l_{ik}\}$ where $l_{ik} \in L$. The association trajectory hypothesis $\Gamma$ is defined as a set of single trajectory hypothesis, i.e. $\Gamma = \{\Gamma_i\}$. The objective of data association is to maximize the posteriori probability of $\Gamma$ given the tracklets set $L$ under the non-overlap constraints [21]:

$$\Gamma^* = \arg\max_{\Gamma} \prod_i P(l_i | \Gamma) \prod_{\Gamma_k \in \Gamma} P(\Gamma_k) \quad \Gamma_i \cap \Gamma_j = \emptyset, \forall i \neq j.$$  

(1)

$P(l_i | \Gamma)$ is the likelihood of tracklet $l_i$. The prior $P(\Gamma_k)$ is modeled as a Markov chain containing transition probabilities $\prod P(l_{ik+1} | l_{ik})$ of all tracklets in $\Gamma_k$ [57].

The transition probability $P(l_i | l_j)$ is computed using probabilities of appearance feature $P_a(l_i \rightarrow l_j)$ and motion feature $P_m(l_i \rightarrow l_j)$. 

$$P(l_i | l_j) = P(l_i \rightarrow l_j) = (P_a(l_i \rightarrow l_j))^{k_1} \cdot (P_m(l_i \rightarrow l_j))^{k_2},$$

(2)

where $k_1$ and $k_2$ are the ratios of two features.

The MAP association model can be solved by a min-cost flow network [18]. The min-cost flow graph is formulated as $G = \{N, E, W\}$, where $N, E, W$ stands for nodes, edges and weights respectively and the weight means the cost of linking the edge. In the graph $G$, there are two nodes $i^\text{enter}$ and $i^\text{exit}$ defined for each tracklet $l_i$. The observation edge $e_i$ from node $i^\text{enter}$ to $i^\text{exit}$ indicates the likelihood of tracklet $l_i$.

The corresponding observation weight $w_i$ is set to the negative logarithm of the likelihood $P(l_i | \Gamma)$. The possible linking relationship between any two tracklets is expressed as a transition edge $e_{ij}$ from node $i^\text{enter}$ to node $j^\text{exit}$. The transition weight $w_{ij}$ is the negative logarithm of the transition probability $P(l_j | l_i)$, as shown as follows,

$$w_{ij} = -\log \frac{P(l_j | l_i)}{1 - P(l_j | l_i)}.$$  

(3)

The transition weight can also be decomposed into probabilities in continuity of appearance and motion,

$$w_{ij} = -\log P(l_j | l_i) = -k_1 \log P_a(l_i \rightarrow l_j) - k_2 \log P_m(l_i \rightarrow l_j).$$  

(4)

In addition to these nodes and edges, there are two extra nodes $S, T$. They are virtual source and sink for the min-cost flow graph. The enter/exit edges $e_{Si}$ and $e_{Ti}$ are also added in to represent the start tracklet $l_i$ and the end tracklet $l_j$. The enter/exit weights of these tracklets are both set to 0 in this paper, because every node could be a start or end tracklet without cost.

In sum, the number of $N$ is $(2M + 2)$, and the numbers of $E$ and $W$ are smaller than the numbers of full connection graph $(3M + 2 \times \binom{2M}{2})$. $M$ is the total number of tracklets in all cameras. As shown in Fig. 3 the graph is solved by the min-cost flow, and the optimal solution is the maximum of the posteriori probability of $\Gamma$ with the minimum cost.

In the rest of this section, we introduce every part of the min-cost flow graph, especially for the weight $W$. 

\section{A. Nodes}

In the proposed approach, the fragmentary tracklets extracted by single-object tracking methods are treated as input observations instead of detections. In other words, these fragmentary tracklets are used as nodes in the global graphic model. One of the reasons is that nodes have more information (like motion) than detections which only contains appearance information. With more information, they can be considered as a credible node and the similarities of them are more reliable. What’s more, as nodes, the tracklet number is much

![Illustration for the min-cost flow network](image-url)

Fig. 3. Illustration for the min-cost flow network. An example for the min-cost flow network with 3 timesteps and 6 tracklets. The number of $N, E$ and $W$ are 14, 21 and 21.
smaller than that of detections. It’s a good way to speed up the computing time of the graph optimization. In this paper, a head-shoulder detector [58] and an AIF tracker [59] are first used to get all the tracklets from each camera. After obtaining detections by the head-shoulder detector, the AIF tracker is applied on tracking every target and getting their tracklets. During the target tracking by the AIF, a confidence $\alpha_t$ [59] is got to evaluate the accuracy of a tracking result in frame $t$. If the confidence score is lower than the threshold, i.e. $\alpha_t < \theta$, 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 $P(l_i|\Gamma)$ of tracklet $i$,
\begin{equation}
 c_i = P(l_i|\Gamma) = \frac{\sum_{k=t_{i,\text{start}}^e}^{t_{i,\text{end}}^e} \alpha_k}{t_{i,\text{end}}^e - t_{i,\text{start}}^e},
\end{equation}
where $t_{i}^{\text{start}}$ and $t_{i}^{\text{end}}$ are the start and end frame for tracklet $i$.

So all the tracklets from all cameras are obtained, $L = \{l_1, l_2, ..., l_M\}$, where each tracklet $l_i = [x_i, c_i, s_i, t_i, \alpha_i]$ consists of position, likelihood, camera view, time stamp and appearance information respectively. The nodes $N$ can be expressed as:
\begin{equation}
 N = \{S, T, l_i^{\text{enter}}, l_i^{\text{exit}}\} i \in [1, M]
\end{equation}

\textbf{B. Edges}

Edges are also an important part for the graph model. All the observation edges and enter/exit edges are reserved in the min-cost flow graph. However, for the transition edges, only a part of it is retained because that not all the edges are meaningful. Three rules are built for selecting transition edges in our graph.

Firstly, for edge $e_{ij}$, the start frame $t_{ij}^s$ of tracklet $l_i$ must be before the end frame $t_{ij}^e$ of tracklet $l_j$ without any overlapped frame. This rule keeps the edges directed and one-way along time. Secondly, the two tracklets $l_i$ and $l_j$ should come from the same camera or two cameras with an existing topological connection, which ensures the link of two tracklets possible from a panoramic view. Thirdly, a waiting time threshold $\eta$ is brought in to limit the link of two tracklets. If the time interval between two tracklets are wide enough, wider than the threshold, the likelihood of this link is close to zero. As a result, the edges following all rules are selected and reserved,
\begin{equation}
 E = \{e_i\} \cup \{e_{Si}, e_{IT}\} \cup \{e_{ij}\} i \in [1, M], 0 < t_{ij}^{\text{start}} - t_{ij}^{\text{end}} < \eta, \text{Topo}(s_i, s_j) = 1,
\end{equation}
where $\text{Topo}(s_i, s_j)$ means the camera views of tracklet $l_i$ and $l_j$ have an existing topological connection.

For all these selected edges $E$, the capacity is set to an integer ranging from 0 to 1, because every target should be at one and only one place in the same time. If the capacity is 1 in the optimal solution, which means this edge exists and the two tracklets of the edge belong to the same target.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{Illustration of computing the periodic time for a tracklet. An example for a tracklet with the length $\sigma$ of 9 frames. The Avg Sim column shows the validity of every possible periodic time $\tau$. The maximum in Avg Sim column indicates the best periodic time $\tau$ for this tracklet.}
\end{figure}

\textbf{C. Weights}

Weight is an essential attribution for links. In this section, we discuss about the cost attribution of links. The cost indicates the relationship between nodes (tracklets), so the similarity among tracklets is a good choice to be chosen as an indicator for cost. As said above, the weight $W$ is consisted of three parts, the same as edges:
\begin{equation}
 W = \{w_i\} \cup \{w_{Si}, w_{IT}\} \cup \{w_{ij}\} i \in [1, M]
\end{equation}

The observation weight can be obtained according to Eq. 5. And the enter/exit weights are all set to 0 as mentioned above.

In the transition weight, the appearance similarity $P_a(l_i \rightarrow l_j)$ and motion similarity $P_m(l_i \rightarrow l_j)$ are used to form the weight. In the following section we introduce both of them respectively.

\begin{equation}
\begin{align*}
 w_i &= -\log \frac{P(l_i|\Gamma)}{1-P(l_i|\Gamma)} = -\log \frac{c_i}{1-c_i}, \\
 w_{Si} &= w_{IT} = 0, \\
 w_{ij} &= -\log P(l_j|l_i) = -k_e \log P_a(l_i \rightarrow l_j) - k_m \log P_m(l_i \rightarrow l_j).
\end{align*}
\end{equation}

\textbf{1) Appearance Similarity:} As shown in Section II, both SCT and ICT have their own representations and similarity metrics, while ICT’s methods are more sophisticated than SCT’s. In order to build an equalised metric, the proposed approach adopts an ICT representation. But, without learning process, this may strongly reduce the computing speed due to so many added links. The representation is called piecewise major color spectrum histogram representation (PMCSHR) [18]. It’s an improved version of major color spectrum histogram representation (MCSHR) [20] by extracting some periodicity information that is specific to pedestrian. MCSHR obtains the major colors of a target based on an online k-means clustering algorithm. When computing the MCSHR for a tracklet, one normal way is to integrate the histograms from all the frames together,
\begin{equation}
 H_i = \sum_{n=1}^{\sigma} h_n^{l_i},
\end{equation}
where $h_{t_i}^i$ is the MCSHR for tracklet $l_i$ in the $n$th frame and $H_i$ is the incremental MCSHR [20] for the whole tracklet $l_i$. $\varpi_i$ is the length of tracklet $l_i$.

As a non-rigid target, Pedestrian is not much like vehicle in tracking, and its MCSHR is still challenge for tracking and identification. However the pedestrian has its own special characteristic. When a person walks, his walking has significant 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 $\tau_i$ and use it to segment tracklets.

As a preparation, all MCSHRs $\{h_1, h_2, ..., h_{\varpi_i}\}$ for tracklet $l_i$ are obtained and the similarity $\Lambda_{k,j}$ between any pair $h_k$ and $h_j$ is computed. The idea is to compute all the possible periodic times and find the best one. For a certain periodic $\tau$ periodically in a certain time. It assumes that people always walk at a constant speed. The proposed solution is to find this periodic time $\tau_i$ and use it to segment tracklets.

As a preparation, all MCSHRs $\{h_1, h_2, ..., h_{\varpi_i}\}$ for tracklet $l_i$ are obtained and the similarity $\Lambda_{k,j}$ between any pair $h_k$ and $h_j$ is computed. The idea is to compute all the possible periodic times and find the best one. For a certain periodic $\tau$ periodically in a certain time. It assumes that people always walk at a constant speed. The proposed solution is to find this periodic time $\tau_i$ and use it to segment tracklets.

$\tau_i = \arg \max_t \frac{1}{\varpi_i - t} \sum_{j=1}^{\varpi_i - t} \Lambda_{j,j+t} \quad \forall t \in [\gamma, \varpi_i/2). \quad (11)$

The set $[\gamma, \varpi_i/2)$ is used to limit the possible range of $t$, and $\gamma$ is set to 15. If $\gamma$ is too small, the nearby frames have a strong similarity which causes Eq. (11) to a false maximum. After calculation, $\tau_i$ is the best periodic time for tracklet $l_i$. Then the tracklet $l_i$ can be evenly segmented into pieces with the length $\tau_i$ (except the end part). For each piece, the incremental MCSHR is computed. The PMCSHR for tracklet $l_i$ is represent as $\{H_1^i, H_2^i, ..., H_{d_i}^i\}$, and $d_i = \lceil \frac{\varpi_i}{\tau_i} \rceil$ is the number of pieces the tracklet $i$ segmented.

Then every similarity between each two pieces from tracklets $l_i$ and $l_j$ are computed, the average $Dis(l_i, l_j)$ of which is considered as the appearance similarity between two tracklets.

$$Dis(l_i, l_j) = \frac{1}{d_i \times d_j} \sum_{n=1}^{d_i} \sum_{m=1}^{d_j} Sim(H_n^i, H_m^j), \quad (12)$$

where $Sim(H_n^i, H_m^j)$ is the similarity metric for two tracklet incremental MCSHRs.

During tracking objects in a single camera, we assume that observations are obtained under the same circumstance, like illumination and angle of view. Hence the targets would have a strong invariance in their appearance representation which can further be used for tracking. During inter-camera object tracking, this invariance is weaker due to the changes in different circumstances. When we establish the graph with nodes and edges, this phenomenon would cause the inter-camera similarity being much lower than the similarity in single camera. If there's no alignment or equalization for two similarity distributions or compensation for the inter-camera similarity, it would indicate the graph links the edges in the single camera preferentially all the time and ignores the inter-camera links as long as a higher edge existing in the same camera, even it may be much smaller compared with others. As it's hard to get an accurate alignment for two similarity distributions, the proposed approach offers a suitable alignment which can be considered as a compensation for the inter-camera similarity. Our purpose is to equalised the difference between two similarity distribution and at the same time try to keep the distribution of the inter-camera similarity not affected. So the equalization is processed on the distribution of the single camera similarity.

$$P_a(l_i \rightarrow l_j) = \frac{\Delta \sigma(Dis(l_i, l_j) - \Delta \mu)}{\Delta \sigma}, \quad (13)$$

where $\Delta \sigma$ and $\Delta \mu$ are the compensation factors, the $Dis(l_i, l_j)$ are obtained by Eq. (12).

The $\Delta \mu$ is used to improve the average level of the single camera similarity distribution and the $\Delta \mu$ is utilized to control the amplitude of variation. They are computed from two similarity distributions.

$$\Delta \mu = \mu_1 - \mu_2, \quad \Delta \sigma = \sigma_2/\sigma_1, \quad (14)$$

where $\mu_1$ and $\sigma_1$ are the mean and variance of the single camera similarity distribution. These should be computed by all the single camera edges. And $\mu_2$ and $\sigma_2$ are of the inter-camera similarity distribution and should be got from all the inter-camera edges.

However, not all the similarities of edges are reliable and suitable. Some have a large proportion of noises and should be excluded as outliers. In this paper, a minimum uncertain gap (MUG) [19] is brought in to help to filtrate edges used for computing the mean and variance. The MUG tries to find the relationship which minimizes uncertainty of the likelihood between two tracklets. The tracklet link with a small MUG can be excluded as outliers. In this paper, a minimum uncertain gap (MUG) [19] is brought in to help to filtrate edges used for computing the mean and variance. The MUG tries to find the relationship which minimizes uncertainty of the likelihood between two tracklets. The tracklet link with a small MUG can be excluded as outliers. In this paper, a minimum uncertain gap (MUG) [19] is brought in to help to filtrate edges used for computing the mean and variance. The MUG tries to find the relationship which minimizes uncertainty of the likelihood between two tracklets. The tracklet link with a small MUG can be excluded as outliers. In this paper, a minimum uncertain gap (MUG) [19] is brought in to help to filtrate edges used for computing the mean and variance. The MUG tries to find the relationship which minimizes uncertainty of the likelihood between two tracklets. The tracklet link with a small MUG can be excluded as outliers.

$$\text{MUG}(l_i, l_j) = \max Sim(H_{n_i}, H_{m_i}) - \min Sim(H_{n_i}^i, H_{m_i}^j), \quad (15)$$

Therefore, with the help of MUG’s filtration, the mean and variance are computed as follows:

$$\mu_1 = \text{MEAN}(Dis(l_i, l_j)), \quad \sigma_1 = \text{VAR}(Dis(l_i, l_j)), \quad \text{MUG}(l_i, l_j) < \varepsilon, s_i = s_j. \quad (16)$$

$$\mu_2 = \text{MEAN}(Dis(l_i, l_j)), \quad \sigma_2 = \text{VAR}(Dis(l_i, l_j)), \quad \text{MUG}(l_i, l_j) < \varepsilon, s_i \neq s_j, \quad (17)$$

where $\varepsilon$ is a confidence threshold, MEAN() and VAR() are the mean and variance operations respectively.

And the final equalised appearance similarity metric is:

$$P_a(l_i \rightarrow l_j) = \left\{ \begin{array}{ll}
Dis(l_i, l_j) & \text{if } s_i \neq s_j, \\
\Delta \sigma(Dis(l_i, l_j) - \Delta \mu) & \text{if } s_i = s_j.
\end{array} \right. \quad (18)$$
2) Motion Similarity: As a general method that is available in both overlapping and non-overlapping views, it’s hard to always build an exact 3D coordinate system to project all scenes together. As a result, in this paper, a relative distance between two tracklets is adopted to measure the motion similarity. For two tracklets \( l_i \) and \( l_j \), it’s easy to get their interval time by a simple subtraction. If the two tracklets are likely to belong to one target, the interval time \( t_{ij}^{\text{inv}} \) must be a positive number.

\[
\begin{align*}
  t_{ij}^{\text{inv}} &= t_{ij}^{\text{start}} - t_{ij}^{\text{end}},
\end{align*}
\]

where \( t_{ij}^{\text{start}} \) is the start time for tracklet \( j \) and \( t_{ij}^{\text{end}} \) is the end time for tracklet \( i \).

With the interval time \( t_{ij}^{\text{inv}} \), the position \( x_i^{t_{ij}^{\text{inv}}} \) of tracklet \( i \) after \( t_{ij}^{\text{inv}} \) time, we can predict the position for tracklet \( i \) after \( t_{ij}^{\text{inv}} \) time. The new position can be calculated below:

\[
\begin{align*}
  x'_i &= x_i^{t_{ij}^{\text{inv}}} + v_i^{t_{ij}^{\text{inv}}} \times t_{ij}^{\text{inv}}.
\end{align*}
\]

As people usually walk along a smooth path in real scenes, we can assume that, if the two tracklets belong to a same person, the corresponding predicted position must be close to each other. In other words, \( x'_i \) and \( x'_j \) should be near enough to \( x_i^{\text{head}} \) and \( x_i^{\text{end}} \) respectively. Therefore, the distances between predicted positions and original positions are used to represent the motion similarity between two tracklets (seen in Fig. 5).

So the motion similarity in the single camera is computed as below:

\[
P_m(l_i \rightarrow l_j) = \exp(-\lambda \frac{1}{2} (\Delta x_i + \Delta x_j)) \quad s_i = s_j.
\]

From Eq. 22 it’s easy to find that the relative distance is only valid for two tracklets from the same camera. If tracklets are from different cameras, the interval time is partly invalid. Because in inter-camera cases, the paths between cameras are hard to measure which causes the interval time useless for predicting positions. In this case, the relative distance mostly tends to be a huge wrong number. To handle this problem, a minimum relative distance is applied to compute the similarity across cameras and make it comparable with Eq. 22.

As we know, enter/exit areas are commonly used in some uncalibrated camera systems to help re-local targets’ exact positions. First, we labeled enter/exit areas for each camera view with the help of topology information (see Fig. 6).

For a person, if she gets disappeared from an exit area, we assumed that she would be found in the enter area of the corresponding camera (seen in Fig. 7 (a)). If she gets disappeared from an area near a exit area, she could re-appear in the corresponding enter area with a high probability. In this case, besides enter/exit areas, we set a disappearing point manually for each area to connect cameras. Then a minimum relative distance \( \Delta x_{i}^{\text{min}} \) to the disappearing point during the whole interval time is adopted to measure the motion similarity across cameras instead of the relative distance \( \Delta x_i \) computed by the predicted position after the interval time, seen in Fig. 7 (b) and (c).

\[
\Delta x_{i}^{\text{min}} = \begin{cases} 
  \min_{t \in [1, t_{ij}^{\text{inv}}]} \| x_i^{t_{ij}^{\text{inv}}} + v_i^{t_{ij}^{\text{inv}}} \times t - x_i^{\text{exit}} \|_2, & \text{if } x_i^{t_{ij}^{\text{inv}}} \notin \text{Area}_{\text{exit}}, \\
  0, & \text{if } x_i^{t_{ij}^{\text{inv}}} \in \text{Area}_{\text{exit}}, 
\end{cases}
\]

As a general method that is available in both overlapping and non-overlapping views, it’s hard to always build an exact 3D coordinate system to project all scenes together. As a result, in this paper, a relative distance between two tracklets is adopted to measure the motion similarity. For two tracklets \( l_i \) and \( l_j \), it’s easy to get their interval time by a simple subtraction. If the two tracklets are likely to belong to one target, the interval time \( t_{ij}^{\text{inv}} \) must be a positive number.

\[
\begin{align*}
  t_{ij}^{\text{inv}} &= t_{ij}^{\text{start}} - t_{ij}^{\text{end}},
\end{align*}
\]

where \( t_{ij}^{\text{start}} \) is the start time for tracklet \( j \) and \( t_{ij}^{\text{end}} \) is the end time for tracklet \( i \).

With the interval time \( t_{ij}^{\text{inv}} \), the position \( x_i^{t_{ij}^{\text{inv}}} \) of tracklet \( i \) after \( t_{ij}^{\text{inv}} \) time, we can predict the position for tracklet \( i \) after \( t_{ij}^{\text{inv}} \) time. The new position can be calculated below:

\[
\begin{align*}
  x'_i &= x_i^{t_{ij}^{\text{inv}}} + v_i^{t_{ij}^{\text{inv}}} \times t_{ij}^{\text{inv}}.
\end{align*}
\]

As people usually walk along a smooth path in real scenes, we can assume that, if the two tracklets belong to a same person, the corresponding predicted position must be close to each other. In other words, \( x'_i \) and \( x'_j \) should be near enough to \( x_i^{\text{head}} \) and \( x_i^{\text{end}} \) respectively. Therefore, the distances between predicted positions and original positions are used to represent the motion similarity between two tracklets (seen in Fig. 5).

So the motion similarity in the single camera is computed as below:

\[
P_m(l_i \rightarrow l_j) = \exp(-\lambda \frac{1}{2} (\Delta x_i + \Delta x_j)) \quad s_i = s_j.
\]

From Eq. 22 it’s easy to find that the relative distance is only valid for two tracklets from the same camera. If tracklets are from different cameras, the interval time is partly invalid. Because in inter-camera cases, the paths between cameras are hard to measure which causes the interval time useless for predicting positions. In this case, the relative distance mostly tends to be a huge wrong number. To handle this problem, a minimum relative distance is applied to compute the similarity across cameras and make it comparable with Eq. 22.

As we know, enter/exit areas are commonly used in some uncalibrated camera systems to help re-local targets’ exact positions. First, we labeled enter/exit areas for each camera view with the help of topology information (see Fig. 6).

For a person, if she gets disappeared from an exit area, we assumed that she would be found in the enter area of the corresponding camera (seen in Fig. 7 (a)). If she gets disappeared from an area near a exit area, she could re-appear in the corresponding enter area with a high probability. In this case, besides enter/exit areas, we set a disappearing point manually for each area to connect cameras. Then a minimum relative distance \( \Delta x_{i}^{\text{min}} \) to the disappearing point during the whole interval time is adopted to measure the motion similarity across cameras instead of the relative distance \( \Delta x_i \) computed by the predicted position after the interval time, seen in Fig. 7 (b) and (c).

\[
\Delta x_{i}^{\text{min}} = \begin{cases} 
  \min_{t \in [1, t_{ij}^{\text{inv}}]} \| x_i^{t_{ij}^{\text{inv}}} + v_i^{t_{ij}^{\text{inv}}} \times t - x_i^{\text{exit}} \|_2, & \text{if } x_i^{t_{ij}^{\text{inv}}} \notin \text{Area}_{\text{exit}}, \\
  0, & \text{if } x_i^{t_{ij}^{\text{inv}}} \in \text{Area}_{\text{exit}}, 
\end{cases}
\]

where \( x_i^{\text{exit}} \) and \( x_i^{\text{enter}} \) are the positions of the disappearing points for enter area and exit area in camera \( s_i \) respectively.
The minimum relative distance has another benefit that this value is measured in each camera which is comparable with the relative distance. With its help, the motion similarity metric can be extend from a single camera to a multi-camera system and can be considered as well equalised in the global graph.

The final equalised motion similarity metric is:

\[
P_m(l_i \rightarrow l_j) = \begin{cases} 
\exp(-\frac{1}{2}(\Delta x_i + \Delta x_j)) & \text{if } s_i = s_j \\
\exp(-\frac{1}{2}(\Delta x_i^{\min} + \Delta x_j^{\min})) & \text{if } s_i \neq s_j,
\end{cases}
\]

where \(\lambda\) is set to 0.01 in the experiments.

IV. EXPERIMENT RESULTS

In this section, the proposed approach are evaluated based on the following aspects. First, the global graphic model is compared with the traditional two-step framework, where we use the same feature representation for fairness. Second, a performance comparison between the equalised graph and the non-equalised one is provided to prove the effectiveness of the equalization process with the new similarity metric. Third, the proposed approach is compared with some state-of-the-art MCT methods. However, as there’m no benchmark for MCT, we introduce a dataset and a comprehensive evaluation criteria for MCT first, which can be developed as a benchmark in further works. The dataset is specialized for multi-camera pedestrian tracking in non-overlapping cameras called NLPR_MCT dataset. The details of the dataset are presented in Section IV-A. The proposed evaluation criteria for MCT is introduced in Section IV-B.

A. Datasets

For comprehensive performance evaluation, it is crucial to develop a representative dataset. There exist several datasets for visual tracking in the surveillance scenarios, such as PETS [60], CAVIAR [61], TUD [62] and i-LIDS [63] databases. However, most of them are developed for multi-object tracking in single camera and are not much suitable for inter-camera object tracking, especially object re-identification. PETS is under a simulation environment with overlapping cameras, not in real scene, while i-LIDS aims to serve multi-camera object tracking indoor and the ground truth is not for free so far. For these reasons, a new pedestrian dataset is constructed in this paper for multi-camera object tracking to facilitate tracking evaluation.

The NLPR_MCT dataset [14] consists of four sub-datasets. Every sub-dataset includes 3-5 cameras with non-overlapping scenes and has a different situation according to the number of people (ranging from 14 to 255) and the level of illumination changes and occlusions. The collected videos contain real scenes and simulation environments. We also list the topological connection matrixes for pedestrian areas. All the videos are nearly 20 minutes (except Dataset 3) with a rate of 25 fps 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. The connection relationships between scenes are shown in Fig. 8, where the enter/exit areas for this paper are also marked.

B. Evaluation Criteria

As we know, each of SCT and ICT has its own evaluation criteria. Most SCT trackers usually use the multi-object track-
ing accuracy (MOTA) and ID switch [64] as evaluation criteria, while some SCT papers prefer other terms [10], [24], [42]. In ICT, ID switch is a necessary term. And as an important part of ICT, object re-identification has its own evaluation criteria, like CMC and nAUC [47], which is hard to be extended to evaluate the ICT performance.

The two criterions mentioned in section I are important to a multi-camera multi-object tracking system. The SCT and the ICT correspond to the two criterions respectively. As these two criterions are equally crucial for multi-camera object tracking performance, they should be considered equally important in the performance measure.

Nevertheless, in existing multi-camera object tracking, there is rarely a widely accepted performance measure that takes these two criterions into account. The common criteria researchers used is an extension of MOTA and adapted it to multi-camera object tracking. It adds the ID switches in SCT and in ICT together, resulting ignoring the different incidence densities of the ID switches in SCT and ICT. Because in most video scenes, i.e., Fig. 9, the ground truthes used for frame matching in SCT are much more than those in ICT. It makes trackers care more about the trajectories in single camera rather than the inter-camera matching. In this paper, we treat them separately and provide a new evaluation criteria to measure the performance of multi-camera object tracking. Our criteria takes both of two criterions into account and uniform them into one evaluation metric. The metric is called multi-camera object tracking accuracy (MCTA):

\[ MCTA = \frac{Detection \times \text{Tracking}^{SCT} \times \text{Tracking}^{ICT}}{2 \times \text{Precision} \times \text{Recall}} \]

\[ = \left( \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \left( 1 - \frac{\sum mme_s^t}{\sum tp^s_t} \right) \left( 1 - \frac{\sum mme_c^t}{\sum tp^c_t} \right) \]

\[ \text{Detection} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

\[ \text{Precision} = 1 - \frac{\sum fp_t}{\sum rt_t} \]

\[ \text{Recall} = 1 - \frac{\sum mf_t}{\sum gt_t} \]

where \(fp_t, rt, mf_t,\) and \(gt_t\) are the number of false positives, results, misses and ground truth respectively present in time \(t\).

\[ \text{Tracking}^{SCT} = 1 - \frac{\sum mme_s^t}{\sum tp^s_t} \]

\[ \text{Tracking}^{ICT} = 1 - \frac{\sum mme_c^t}{\sum tp^c_t} \]

For SCT and ICT ability parts, we measure the abilities via the number of mismatches (ID-switches). We split the number of mismatches \(mme_t\) in MOTA [64] into \(mme_s^t\) and \(mme_c^t\), \(mme_s^t\) represents the number of mismatches happened in a single camera and \(mme_c^t\) is for those across-camera mismatches. The \(tp_s^t\) and \(tp_c^t\) shares similar meanings, \(tp_s^t\) is the number of ground truth whose last frame and current frame
are in the same camera, while \( tp_c \) means the number of those across-camera ground truth. For a new target, it’s counted as an across-camera ground truth by default in our criteria.

C. Global Graph Model vs Two-Step Framework

The advantage of the proposed solution is to improve the ICT performance under an unperfect SCT result. So in this section, the proposed global graph model is compared with the traditional two-step framework, a SCT approach plus a ICT approach. We use the same MAP model to solve the data association in both SCT and ICT processes of the two-step framework and aim to remove the interference of different data association methods. Adopting the MAP model in SCT is presented in Zhang et al. [21]. However using MAP model in ICT is not a suitable solution when the tracking results in single camera are perfect and unchangeable. But as we said in Section. II-B when the SCT results are not ideal, the data association in ICT should be more like a global optimization problem than the K-partite graph matching problem. The MAP model can be solved by a global optimal solution. That’s another reason why we use the MAP model to achieve the data association of ICT. It should be noticed that this global optimization is to solve the data association in ICT while the proposed equalised global graphic model is for the whole MCT. As a complement, we also utilize Hungary algorithm [46] to achieve the ICT step. Hungary graph is a classical data association method for ICT. The feature representation in this experiment is the same PMCSHR appearance and motion features due to the fairness reason.

In this experiment, the waiting time threshold \( \eta \) and the minimum value \( \varepsilon \) of the MUG are set to 60*25*1 and 0.4 respectively, the ratios of two features \( k_1 \) and \( k_2 \) are both to 1. To prove the ability of the proposed approach handling fragmental tracklets in SCT, the experiment changes the threshold \( \theta \) of the confidence of AIF tracker to produce more fragments artificially. The threshold \( \theta \) ranges from 0 to 0.2 and the corresponding numbers of tracklets are listed beside it.

The total single-camera ground truth number \( tp_s \) and across-camera ground truth number \( tp_c \) for each sub-dataset are listed in Fig. 2. From the first two row in Fig. 10 we can see that with the increase of the tracklets, both the single camera mismatch number \( mme^s \) and the across-camera mismatch number \( mme^c \) grow significantly in the proposed global graph and the two-step framework. In the first row, the single camera mismatch number \( mme^s \) in the proposed global graph is always larger than that in the two-step framework [21], because the two-step framework offers an optimization in each camera which makes it have a better local result. In dataset 3 and dataset 4, the \( mme^s \) in the proposed global graph becomes comparable with that in the two-step framework [21]. The reason is that these two datasets are under simulation condition which has many frequent “walking around” behaviors. On this occasion, the across-camera information may provide more useful feedbacks for each specific camera and can partly improve the SCT performance. For the across-camera mismatch number \( mme^c \) in the middle row, the number in

|                | Dataset1 | Dataset2 | Dataset3 | Dataset4 | AverageMCTA |
|----------------|----------|----------|----------|----------|--------------|
| NonA \( mme^s \) | 71       | 83       | 59       | 125      | 0.3845       |
| NonA \( mme^c \) | 76       | 109      | 71       | 137      | 0.4769       |
| M \( mme^s \)   | 123      | 201      | 132      | 187      | 0.2687       |
| M \( mme^c \)   | 88       | 164      | 116      | 169      | 0.3388       |
| EqlA \( mme^s \) | 101      | 126      | 95       | 188      | 0.2649       |
| EqlA \( mme^c \) | 49       | 107      | 80       | 159      | 0.5066       |
| EqlA+M \( mme^s \) | 66      | 73      | 51       | 128      | 0.5069       |
| EqlA+M \( mme^c \) | 49       | 107      | 80       | 159      | 0.6099       |
| MCTA            | 0.6311   | 0.201    | 0.1312   | 0.2687   | 0.3845       |
|                | 0.7357   | 0.5973   | 0.2359   | 0.3388   | 0.4769       |
|                | 0.6971   | 0.6907   | 0.3735   | 0.2649   | 0.5066       |
|                | 0.8525   | 0.7370   | 0.4724   | 0.3778   | 0.6099       |

Fig. 11. Empirical comparison of the proposed approach on four multi-camera tracking datasets. The experiment is run on all four datasets. NonA is the tracker with non-equalised appearance feature, EqlA is the equalised one. M is corresponding to the tracker with motion feature only and EqlA+M is that combines equalised appearance feature and motion feature together. The bold indicates the best performance. Our proposed method with equalised appearance and motion features always keep the best. It’s also clear that although the equalised-appearance tracker produces more mismatches than the non-equalised tracker during the single camera tracking, in the inter-camera tracking, it always holds a smaller mismatch number than the non-equalised one and has a better MCTA score.

D. Equalised vs Non-equalised Graph Model

This experiment is conducted to prove the effectiveness of the similarity equalization process. All the trackers are under our global graph model. We compare the equalised appearance similarity metric with the non-equalised one and then combined with our equalised motion metric. Particularly, in this experiment, the confidence threshold \( \theta \) of AIF tracker is fixed and set to 0.

The results are shown on Fig. 11. It can be found that the tracker with non-equalised appearance similarity has a lower
mismatch number $mme^s$ in the single camera compared with equalised one. It means that when we conduct equalization, the single camera performance drops down due to the changes of the distribution of the single camera similarity, that is unavoidable but acceptable. In the inter-camera tracking, it is clear that the equalised appearance similarity tracker gives a great help to reduce the number $mme^c$ of mismatches across cameras. When the equalised motion information is added in, the $mme^c$ further decreases. The MCTA is the final comprehensive score which takes both SCT and ICT performances into account. The larger the score is, the better performance the tracker has. As seen in Fig. 11, the equalised appearance similarity tracker combined with the equalised motion information has a highest score. It indicates that the increased single camera mismatch number $mme^s$ in our method is acceptable in order to reduce the across-camera mismatch number $mme^c$ and get a higher score in the whole MCT performance. What’s more, when we use the motion feature alone for the multi-camera object tracking, the performance is comparable and sometimes better than the appearance feature, which partly proves the effectiveness of our equalised motion similarity metrics.

E. Equalised Global Graph Model vs State of The Arts

In this section, we compare our equalised global MAP graph model with other multi-camera object tracking methods. As a comparison, the methods must contain the abilities to handle
both the SCT and ICT steps. We compare the proposed graph with current two-step multi-camera object tracking methods. The methods are from the Multi-Camera Object Tracking (MCT) Challenge [53] in ECCV 2014 visual surveillance and re-identification workshop. USC-Vision ([32], [41]) is the
winner in the challenge which is considered as the state-of-the-art two-step multi-camera object tracking approach. We first conduct the comparison under the condition that the ground truth of single camera object tracking is available, the results are shown in Fig. [2]. It reflects the ICT power of each method when the single camera object tracking results are perfect. From the average MCTA score we can see that USC-Vision \(^{(32, 41)}\) is much better than our proposed method. This proves the advantage of USC-Vision’s ICT method. In Fig. \[13\] only the ground truth of object detection is available, the tracker should achieve the single camera object tracking by themselves. In this case, the SCT results would not be perfect, this step may bring in much fragments and false positive tracklets. From Fig. \[13\] although the SCT performance \(^{\text{mme}^8}\) of USC-Vision \(^{(32, 41)}\) is also better than ours, it is clear that the number of its ICT mismatches increases much more shapely than our method, which indicates that its powerful ICT method loses its advantage under the unperfect SCT results. Some results are shown in Fig. \[14\] As the final evaluation, our equalised global graph model has the highest average MCTA score, which further prove the advantage of our proposed model to improve the ICT performance under an unperfect SCT result.

V. Conclusion

The multi-camera non-overlapping object tracking is an important but challenge task in intelligent visual surveillance systems. We developed a joint approach that integrates the single camera object tracking (SCT) and the inter-camera object tracking (ICT) into one graph. This joint approach overcomes the disadvantages in traditional two-step tracking approaches. In addition, the global graph in the proposed approach is equalised on the similarity metrics, both appearance and motion features. The equalization further reduces the number of mismatch errors in inter-camera object tracking. The results show its effectiveness for multi-camera object tracking, especially when the SCT performance is not perfect. Our approach focuses on the graph modeling instead of feature representation learning. Any existing re-identification feature representation method can be incorporated into our framework.

References

[1] R. Vezzani, D. Baltieri, and R. Cucchiara, “People reidentification in surveillance and forensics: A survey,” ACM Comput. Surv., vol. 46, no. 29, 2013.
[2] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, “Visual tracking: An experimental survey,” IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), vol. 36, no. 7, pp. 1442–1468, 2014.
[3] Y. Pang and H. Ling, “Finding the best from the second best c inhibiting subjective bias in evaluation of visual tracking algorithms,” in IEEE International Conference on Computer Vision (ICCV), 2013.
[4] Y. Wu, J. Liu, and M.-H. Yang, “Online object tracking: A benchmark,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[5] K. Huang and T. Tan, “Vs-star: a visual interpretation system for visual surveillance,” Pattern Recognition Letters (PRL), pp. 2265–2285, 2010.
[6] P. L. Venetianer and H. Deng, “Performance evaluation of an intelligent video surveillance system - a case study,” Computer Vision and Image Understanding (CVIU), vol. 114, no. 11, pp. 1292–1302, 2010.
[7] X. Wang, “Intelligent multi-camera video surveillance: A review,” Pattern Recognition Letters (PRL), vol. 34, pp. 3–19, January 2012.
[8] J. Liu, P. Carr, R. T. Collins, and Y. Liu, “Tracking sports players with context-conditioned motion models,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013.
[9] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. J. V. Gool, “Online multiperson tracking-by-detection from a single, unc�피터ized camera,” IEEE Trans. Pattern Anal. Mach. Intell. (PAMI), vol. 33, no. 9, pp. 1820–1833, 2011.
[10] C.-H. Kuo and N. Nevatia, “How does person identity recognition help multi-person tracking?” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1217–1224.
[11] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, “Tracking across multiple cameras with disjoint views,” in IEEE International Conference on Computer Vision (ICCV), 2003, pp. 952–957.
[12] R. Hamid, R. Kumar, M. Grandmann, K. Kim, I. Essa, and J. Hodgins, “Player localization using multiple static cameras for sports visualization,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2010, pp. 731–738.
[13] Z. Wu, N. Hristov, T. Hedrick, T. Kunz, and M. Betke, “Tracking a large number of objects from multiple views,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1546–1553.
[14] “Nlpr_mct dataset,” http://mct.idealtest.org/Datasets.html.
[15] A. Segal and I. Reid, “Latent data association: Bayesian model selection for multi-target tracking,” in IEEE International Conference on Computer Vision (ICCV), 2013, pp. 2904–2911.
[16] C. Arora and A. Globerson, “Higher order matching for consistent multiple target tracking,” in IEEE International Conference on Computer Vision (ICCV), 2013, pp. 177–184.
[17] A. Butt and R. Collins, “Multi-target tracking by lagrangian relaxation to min-cost network flow,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1846–1853.
[18] W. Chen, L. Cao, X. Chen, and K. Huang, “A novel solution for multi-camera object tracking,” in IEEE International Conference on Image Processing (ICIP), 2014, pp. 3239–2333.
[19] J. Kwon and K. Lee, “Minimum uncertainty gap for robust visual tracking,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2013, pp. 2355–2362.
[20] M. Piccardi and E. Cheng, “Multi-frame moving object track matching based on an incremental major color spectrum histogram matching algorithm,” in IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2005, p. 19.
[21] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–8.
[22] X. Chen, Z. Qin, L. An, and B. Bhanu, “An online learned elementary grouping model for multi-target tracking,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1242–1249.
[23] A. Zamir, A. Dehghan, and M. Shah, “Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs,” in European Conference on Computer Vision (ECCV), 2012, pp. 343–356.
[24] Y. Li, C. Huang, and R. Nevatia, “Learning to associate: Hybridboosted multi-target tracker for crowded scene,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 2953–2960.
[25] M. Yang, Y. Liu, L. Wen, Z. You, and S. Li, “A probabilistic framework for multitarget tracking with mutual occlusions,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[26] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, “Occulsion geodesics for online multi-object tracking,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[27] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele, “Learning people detectors for tracking in crowded scenes,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1049–1056.
[28] C. Dicle, O. Camps, and M. Szaizawi, “The way they move: Tracking multiple targets with similar appearance,” in IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1234–1241.
[29] S. Bae and K. Yoon, “Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
[30] B. Wang, G. Wang, K. Chan, and L. Wang, “Tracklet association with online target-specific metric learning,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1234–1241.
[31] L. Wen, W. Li, J. Yan, Z. Lei, and S. Yi, D. amd Li, “Multiple target tracking based on undirected hierarchical relation hypergraph,” in IEEE
