EURASIP Journal on Advances
in Signal Processing

REVIEW Open Access

Human tracking over camera networks: a review

Li Hou[1,2,3*], Wanggen Wan[1,3], Jenq-Neng Hwang[4], Rizwan Muhammad[1,3], Mingyang Yang[1,3] and Kang Han[1,3]

Abstract

In recent years, automated human tracking over camera networks is getting essential for video surveillance. The tasks of tracking human over camera networks are not only inherently challenging due to changing human appearance, but also have enormous potentials for a wide range of practical applications, ranging from security surveillance to retail and health care. This review paper surveys the most widely used techniques and recent advances for human tracking over camera networks. Two important functional modules for the human tracking over camera networks are addressed, including human tracking within a camera and human tracking across non-overlapping cameras. The core techniques of human tracking within a camera are discussed based on two aspects, i.e., generative trackers and discriminative trackers. The core techniques of human tracking across non-overlapping cameras are then discussed based on the aspects of human re-identification, camera-link model-based tracking and graph model-based tracking. Our survey aims to address existing problems, challenges, and future research directions based on the analyses of the current progress made toward human tracking techniques over camera networks.

Keywords: Human tracking, Generative trackers, Discriminative trackers, Human re-identification, Camera-link model-based tracking, Graph model-based tracking

1 Review

1.1 Introduction

Nowadays, the growing demand of video surveillance systems in some applications such as public security, transportation control, defense, military, urban planning, and business information statistics has attracted increasing attention, and a large number of networked video surveillance systems are getting installed in public places, for instance, airports, subways, railway stations, highways, parking lots, banks, schools, shopping malls, and military areas. These video surveillance systems not only effectively protect the security of public facilities and citizens, but also seamlessly help to transform to smart city, which has attracted more and more scientific researchers to invest huge funds in research related to intelligent video surveillance. It is observed that the main focus of the current research on intelligent video surveillance mainly lies on video object detection/tracking, and video object activity analysis/recognition. The video object tracking is not only one of the most important techniques in intelligent video surveillance, but also the base of high-level video processing and applications such as the subsequent video object activity analysis and recognition. However, in the video object tracking, human tracking is the most challenging since human may vary greatly in appearance on account of changes in illumination and viewpoint, background clutter, occlusion, non-rigid deformations, intra-class variability in shape and pose. Human tracking includes human tracking within a camera and human tracking across multiple cameras. When a person enters into the field of view (FOV) of a camera, human tracking within a camera is needed. However, when he/she leaves the FOV, the human information is no longer available, thus the limited FOV of a camera cannot meet the needs of wide-area human tracking. In order to widen the FOV, human tracking across multiple cameras has to be used since video streams across multiple cameras covering a wider range of areas, which helps to analyze global activities in the real world. Tracking human across multiple

* Correspondence: houli@shu.edu.cn
[1]School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[2]School of Information Engineering, Huangshan University, 245041 Huangshan, China
Full list of author information is available at the end of the article

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

cameras includes two different scenarios, i.e., overlapping camera views and non-overlapping camera views. In the overlapping camera views' scenario, there is a common FOV area between two cameras' views, and human located in the common area (as shown in the area between cameras 1 and 2 in Fig. 1) will appear simultaneously in both cameras' views. In the non-overlapping camera views' scenario, there is not a common FOV area between two cameras' views, i.e., every camera's view is completely disjointed, and human cannot be seen in the so-called blind area (as shown in the area between cameras 2 and 3 in Fig. 1). Compared with human tracking across overlapping cameras, human tracking across non-overlapping cameras will be more challenging and practical. As a result, human tracking over camera networks is necessary and quite challenging in the intelligent video surveillance.

Many issues have made human tracking over camera networks very challenging, including real-time human tracking, variable number of human tracking, and changing human appearance caused by several complicated attributes such as illumination variation, occlusion, nonrigid shape deformation, background clutters, pose variation within a camera, and dramatically varying human appearance due to greatly changing illuminations, viewpoints, and intra-class variability in shape and pose across non-overlapping cameras. In order to deal with the above challenges during human tracking over camera networks, numerous researchers have proposed a variety of tracking approaches. Different approaches focus on solving different issues in human tracking over camera networks. Typically, they attempt to answer the following questions:

- What should be tracked such as bounding box, ellipse, articulation block, and contour?
- What visual features and their pros/cons are robust and suitable for various human tracking tasks?
- Which kinds of statistical learning approaches and the associated properties are appropriate for human tracking?

Although there are some well-known surveys [1–3] in terms of object tracking. However, existing surveys mainly focus on object tracking within a camera. In this survey, we focus on human tracking over camera networks. The main contributions of this survey are as follows:

1) We divide human tracking over camera networks into two inter-related modules: human tracking within a camera and human tracking across non-overlapping cameras.
2) We review the literatures of human tracking within a camera based on the correlation among the human objects. Specifically, we hierarchically categorize the human tracking approaches within a camera into generative trackers and discriminative trackers.
3) We review the literatures of human tracking across non-overlapping cameras from human objects' matching viewpoint. Specifically, we hierarchically categorize the human tracking across non-overlapping cameras into human re-identification (re-id), camera-link model (CLM)-based tracking and graph model (GM)-based tracking.

The rest of the paper is organized as follows: Section 2 gives an overview of the taxonomy of human tracking. Section 3 reviews some core techniques for human tracking within a camera. Section 4 reviews some core techniques for human tracking across non-overlapping cameras, followed by the Conclusions in Section 5.

2 Taxonomy of human tracking

Figure 2 shows the taxonomy of human tracking over camera networks, which is composed of two crucial

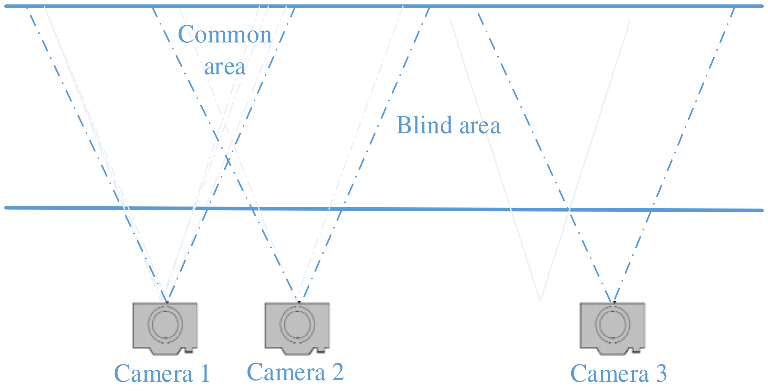


Fig. 1 An example for the topology of a camera network

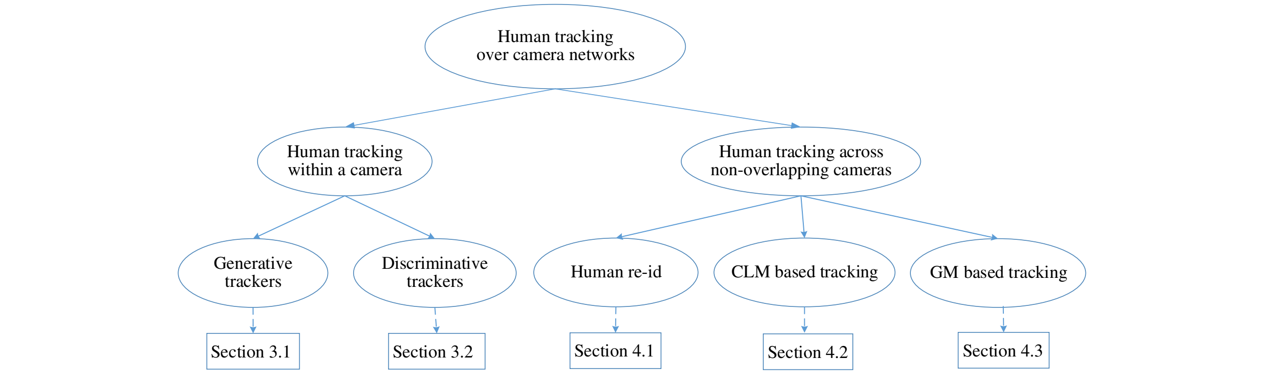


Fig. 2 The taxonomy of human tracking over camera networks

functional modules, i.e., human tracking within a camera and human tracking across non-overlapping cameras. The human tracking within a camera focuses on locating human objects in each frame of a given video sequence from a camera, while the human tracking across non-overlapping cameras concentrates on associating one tracked human object from the FOV of a camera with that from the FOV of another camera. Figure 3 shows the inter-relation between both functional modules.

In the human tracking module within a camera, two kinds of tracking methods including generative trackers and discriminative trackers are discussed, as illustrated by the tree-structured taxonomy in the left part of Fig. 2. The generative trackers focus on searching the most similar target candidate with the minimal reconstruction error in each video frame, while the discriminative trackers aim to separate targets from the background through a classifier, and then to associate the targets frame-by-frame. For a clear illustration of this module, a more detailed literature review of human tracking within a camera is given in Section 3.

As shown in the right part of Fig. 2, the human tracking module across non-overlapping cameras includes three types of tracking methods, i.e., human re-id, CLM-based tracking and GM-based tracking. The human re-id focuses on using visual features of a human object to match with those of the other human objects from different cameras' FOVs based on distance metrics. The spatial (geometrical) relationship, e.g., how far away between a pair of cameras, of these cameras is usually not considered in the process of human re-id. While the CLM-based tracking concentrates on adopting available training data in the corresponding entry/exit zone of two adjacent or multiple neighboring cameras' views to estimate features' mapping relationship (i.e., temporal-spatial relationship and appearance relationship) called a CLM, which can be applied to compensating for the feature differences before computing the feature distance between both human objects across non-overlapping cameras. From the perspective of the optimization framework, the GM-based tracking aims to addressing data association across cameras, which can be modeled as a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG), based on human appearance and motion features. For a clear illustration of this module, a detailed literature review of human tracking within a camera is given in Section 4.

3 Human tracking within a camera

Human tracking within a camera generates the moving trajectories of human objects over time by locating their positions in each frame of a given video sequence. Based on the correlation among the human objects, the human

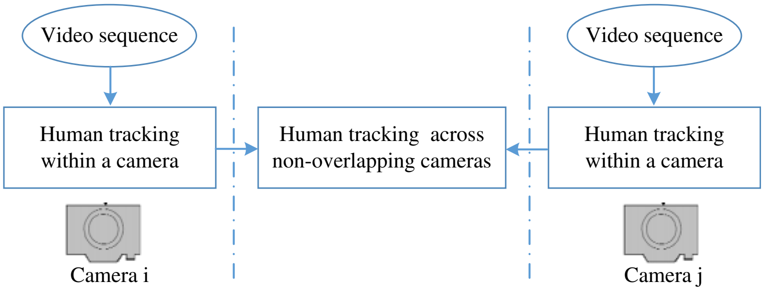

Video sequence
Human tracking within a camera
Camera i
Human tracking across non-overlapping cameras
Video sequence
Human tracking within a camera
Camera j

Fig. 3 The inter-relation between functional modules of human tracking over camera networks

tracking within a camera can be categorized as two types, the generative trackers and the discriminative trackers.

For the generative trackers, each target location and correspondence are estimated by iteratively updating respective location obtained from the previous frame. During the iterative search process for human objects, in order to avoid exhaustive search of the new target location to reduce the cost of computation, the most widely used tracking methods include Kalman filtering (KF) [4–7], Particle filtering (PF) [8–11], and kernel-based tracking (KT) [12–16]. KF expresses a target movement as a dynamic process over the temporal frames and uses the previous target state to predict the next location (and possible size), and then uses the current observation to update the target location. KF can be widely applied to linear/Gaussian real-time tracking. However, when the target state variables do not follow the linear state transition and measurement relationship with Gaussian noise distributions, the KF will give poor state variable estimation results. Moreover, this tracking method cannot deal with target occlusion problem. PF realizes recursive Bayesian filtering through sequential Monte Carlo sampling based on particle representations of probability densities with associated weights. Since the PF generalizes the traditional KF and can be applied to solving non-linear/non-Gaussian tracking problems, it has a wider range of applications due to the superiority in the non-linear and non-Gaussian conditions as well as the multi-modal processing ability. However, PF has relatively high computational complexity, resulting in difficulty in achieving real-time tracking. KT adopts the mean shift (a gradient descent search based optimization method to find local optimal solution) search procedure to find the target candidate which has the highest similarity to the target model, that is represented by a spatially weighted color histogram. KT has gained more popularity for its fast convergence speed and low computation requirement, and thus can achieve real-time tracking. However, when a target is occluded, the conventional KT tends to lose the tracked target because of mismatch between target model and target candidate. Multiple-kernel tracking (MKT) can help to solve the target occlusion problem. The MKT, which extends the conventional KT through representing the tracked target model with multiple kernels, e.g., two kernels (a kernel is expressed as an ellipse) are used to represent the upper/lower half of the human body separately, as shown in Fig. 4. When the lower half of the human body is occluded (left of Fig. 4), using the kernel histogram of the visible upper half of the human body as the target model (right of Fig. 4), the robust human tracking under occlusion can thus be achieved [13]. In order to track the objects more effectively, some constraints among kernels need be considered in the MKT.

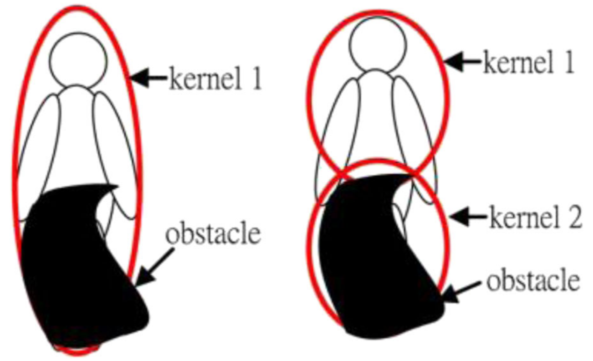


Fig. 4 *Red ellipses* represent kernels. Single kernel (*left*) and two kernels (*right*) with occlusion [13]

While for the discriminative trackers, all the human locations in each video frame are first obtained through a human detection algorithm [17], and then the tracker jointly establishes these human objects' correspondences across frames through a target association technique. The most widely used target association techniques include joint probability data association filtering (JPDAF) [18–21], multiple-hypothesis tracking (MHT) [22–25], and flow network framework (FNF) [26–29]. The JPDAF computes a Bayesian estimate of the correspondence between two consecutive frames, based on calculating all possible target-measurement association probabilities jointly. However, JPDAF only applied to performing data association between a fixed number of tracked targets, otherwise the tracking accuracy will be significantly degraded. The MHT overcomes this limitation by attempting to track all of possible associated hypothesis over several temporal frames and then to determine the most likely target correspondences in the several detected observations. More specifically, the MHT performs data association through building a tree of potential track hypotheses for each candidate target, where the likelihood of each track needs be calculated, and the most likely combination of tracks is selected as the finalized measurement association. However, with the increase in the number of associated objects, its computational cost will increase exponentially. The FNF formulates the target association problem as a minimum cost flow network problem with global optimization for all of the target trajectories. More specifically, the FNF represents the number of targets in the video/image as the amount of flow in the network, while the number of targets is unknown in advance. The goal of the FNF is to globally search for the amount of flow that produces the minimum cost. FNF can effectively achieve multi-target tracking. However, when there are a large number of associated objects, it needs a very high computational cost. Table 1 shows the list of the human tracking algorithms within a camera.

Table 1 Human tracking algorithms within a camera

Method	Description	Typical techniques	Pros	Cons
Generative trackers	To estimate each target's location and correspondence through searching the most similar target candidate with the minimal reconstruction error	KF	Real-time tracking	Subject to linear target state transition and Gaussian noise distributions; apt to lose the tracked target when a target is occluded
		PF	Non-linear/non-Gaussian tracking and multi-modal processing	High computational complexity
		KT	Real-time tracking	Cannot deal with long-term total target occlusion
Discriminative trackers	To separate targets from the background through a classifier, and then jointly to establish these targets' correspondences across frames through a target association algorithm	JPDAF	Multi-target tracking	Subject to data association between a fixed number of tracked targets
		MHT	Variable number of multi-target tracking under occlusion	Vitally high computational requirement
		FNF	Variable number of multi-target tracking under occlusion	Cannot effectively deal with long-time target occlusion

3.1 Generative trackers

Generative trackers are widely applied to human tracking within a camera. Based on different iterative search method for human objects, the generative trackers can be divided into three types, i.e., KF, PF, and KT. Table 2 lists qualitative comparison of generative trackers-based human tracking within a camera.

3.1.1 KF

KF, which has been widely used for tracking problems, can be utilized to predict target motion information to reduce the search area of moving objects. Jang et al. [4] propose active models-based KF tracking algorithm to handle inter-frame changes of non-rigid human objects such as illumination changes and shape deformation. This method applies the framework of energy minimization to active models which characterizes structural and regional features of a human object such as edge, shape, color as well as texture, and hence, adapts dynamically the changes of non-rigid human objects in the consecutive video frames. Moreover, the proposed algorithm adopts KF to predict human objects' motion information to reduce the search space during the human matching process. However, the proposed approach is not applicable to track human objects in occlusion. Jang et al. [5] further propose structural KF to handle objects' occlusion during the human tracking. The proposed algorithm uses relational information of objects' sub-regions to compensate the unreliable measurements of occluded sub-regions. More specifically, the structural KF is composed of two kinds of KFs: cell KF and relation KF. The cell KF estimates motion information of each sub-region of a human body, and the relation KF estimates the relative relationship between two adjacent sub-regions. The final estimation of a sub-region is obtained through combining the involved KFs' estimations.

Table 2 Qualitative comparison of generative trackers-based human tracking within a camera

Item no.	Used generative trackers	Speed	Occlusion	Scale change	Shape deformation
1	Active models-based KF (Jang et al. [4])	High	×	√	√
2	Structural KF (Jang et al. [5])	Moderate	√	√	√
3	adaptive KF (Weng et al. [6])	High	√	√	√
4	features-based KF (Li et al. [7])	Moderate	√	√	√
5	MCMC-based PF (Chong et al. [8])	Low	√	√	√
6	Swarm intelligence-based PF (Zhang et al. [9])	Moderate	×	√	√
7	Occlusion-aware PF (Meshgi et al. [10])	Low	√	√	√
8	interactive PF (Yang et al. [11])	Low	√	√	√
9	Eigenshape-based KT (Liu et al. [12])	High	√	√	√
10	adaptive MKT (Chu et al. [13])	High	√	√	√
11	fragments-based MKT (Fang et al. [14])	High	√	√	√
12	deformable MKT (Hou et al. [15, 16])	Moderate	√	√	√

Symbols √ and× mean that the used generative trackers-based human tracking within a camera can or cannot deal with the situations of occlusion, scale change, and shape deformation

However, the proposed approach is difficult to select a criterion to partition human objects' sub-regions, especially when tracking multiple human objects. Moreover, it needs the other mechanism to judge each human object's degree of occlusion, resulting in a very complex human tracking system. To overcome this drawback, Weng et al. [6] propose a real-time and robust human tracking algorithm in a real-world environment, such as occlusion, lighting changes, fast moving human object, etc., based on adaptive KF, which allows the parameter estimations of KF to adjust automatically. More specifically, the proposed algorithm constructs a motion model to build the system state, which is then applied to prediction step, and uses color features in HSI color space to detect the moving human object so as to obtain the system measurement, where occlusion ratio is used to adaptively adjust the error covariance of KF. Li et al. [7] propose a multi-target (i.e., moving human/vehicle) tracking algorithm using a KF motion model, based on features including the center of mass and tracking window of moving targets. More specifically, the proposed algorithm uses the background subtraction method to detect and extract moving objects, and then the detection results are used to determine whether there is a merge/split among targets. When targets' regions have merged together, multiple moving targets are regarded as a whole target to track for the moment, while when splitting multiple moving targets, feature matching is used to establish corresponding relationship of multiple merged targets, such an example of tracking three human targets in an outdoor scene is shown in Fig. 5. In short, the KF-based tracking algorithm can effectively track objects, but it is only applicable for linear/Gaussian tracking problems.

3.1.2 PF

PF, which generalizes the traditional KF, can be applied to non-linear/non-Gaussian tracking problems. The Markov Chain Monte Carlo (MCMC) method, which samples from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution, is well applied to tracking problems to overcome the limitation of important sampling of original PF in high dimensional state space. Cong et al. [8] propose a robust MCMC-based PF tracking framework, which combines a color-based observation model with detection confidence density derived from histograms of oriented gradients (HOG) descriptor, and adopts MCMC-based particle algorithm to estimate the posterior distribution of the state of a human object to solve the robust human tracking problem. To further handle sample impoverishment problem suffered by conventional PF, Zhang et al. [9] propose a swarm intelligence-based PF tracking algorithm, where particles are firstly propagated through the state transition model, and then corporately evolved according to particle swarm optimization (PSO) iterations based on the cognitive and social aspects of particle populations. The proposed algorithm regards particles as intelligent individuals, and these particles evolve by communicating and cooperating with each other. In this way, the newest observations are gradually considered to approximate the sampling results from the optimal proposal distribution and hence overcome the sample impoverishment problem suffered by conventional PF. To deal with the challenging occlusion problem during human tracking, Meshgi et al. [10] propose an occlusion-aware particle filter framework to deal with complex and persistent occlusions during human tracking. More specifically, the proposed method adopts a binary occlusion flag attached to each particle and treats occlusions in a probabilistic manner. The "occlusion flag" signals whether the corresponding bounding box is occluded, and then triggers the stochastic mechanism to enlarge the objects' search area to accommodate possible trajectory changes during occlusions, meanwhile stops the template updating to prevent the model from being corrupted by irrelevant data. Yang et al. [11] propose interactive PF with

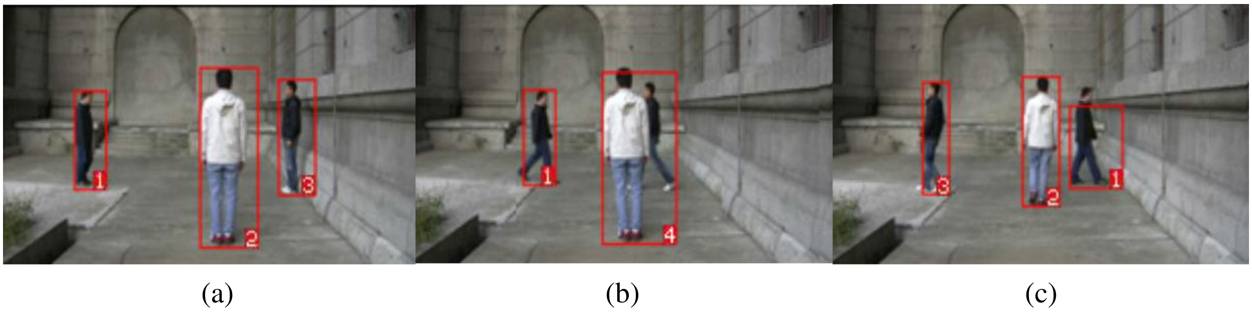


Fig. 5 Illustration of tracking three human targets in an outdoor scene. **a** Three human targets tracked with *red rectangle* bounding boxes labeled 1, 2, and 3. **b** Target labeled 2 and target labeled 3 merge together as a new target labeled 4. **c** Merged targets split into target labeled 2 and 3 by feature matching [7]

occlusion handing for multi-person tracking. More specifically, they use RGB color space model of each human object obtained by human detection operation, and then use the PF on each human object. Further, the proposed algorithm adopts a particle location conflict set to judge the occlusion relationship between different human objects, and chooses the right appearance model adaptively for similarity measurement to update the corresponding particle weights, thus successfully resolves a fully mutual occlusion problem when tracking multiple pedestrians, such an example of tracking multiple human targets in an outdoor scene is shown in Fig. 6. In short, the PF-based tracking algorithms can effectively track the moving human objects, applicable to both linear/Gaussian and non-linear/non-Gaussian tracking problems. However, it requires matching a large number of particles to approximate the posterior probability distribution of the target state, hence it is not applicable to real-time object tracking.

3.1.3 KT

KT has been widely used for real-time target tracking problems. During the target tracking, when a target is moving toward or away from a camera, the scale of the target often changes over temporal frames. In order to overcome the problem, by taking the merit from asymmetric kernel template, Liu et al. [12] propose an eigenshape kernel-based mean shift tracking algorithm to handle the scale changes of tracked objects. The so-called eigenshape kernel refers to an adaptively changing kernel shape by depending on the projection of each tracking window into an eigenshape space. The proposed algorithm utilizes the eigenshape representation, which is obtained by using a principle component analysis method, to construct an arbitrarily shaped kernel so as to adapt to object shape. By making the best of positive correlation between the target size and the corresponding kernel bandwidth, Chu et al. [13] adopt the gradient of the density estimator with respect to the kernel bandwidth to update the scale of tracked objects. The proposed scale-updating method is a simple and effective solution to deal with the target scale change issue. In addition, a target often suffers from the occlusion during target tracking, especially in crowd scenes; it is very difficult for the KT to robustly track the target since single kernel is insufficient to represent the target. To overcome this drawback, MKT is thus proposed in recent years [13–16]. Fang et al. [14] propose MKT based on fragments to deal with occlusion issue. The tracked target is divided into several fragments by integrating the log-likelihood ratio image and morphological operation, and each fragment is tracked through a kernel using the mean shift procedure. Further, to make the best of the inter-relationship among kernels that can provide useful information for tracking, Chu et al. [13] propose adaptive MKT based on the projected gradient optimization algorithm, which combines the total cost function with the constraint functions that defined the inter-relationship among kernels, and hence enables multiple kernels that represents different human body parts to find the best match of the tracked human objects under predefined geometric constraints. However, arbitrary kernel partitioning makes it difficult to define effective geometric constraints among kernels. To better deal with this issue to improve the robustness and effectiveness under occlusion further, Hou et al. [15, 16] propose a deformable multiple-kernel-based human tracking system using a moving camera. This system regards each part model of a deformable part model (DPM) detected human [30] as a kernel, where the DPM represents a human object by a so-called star model, that is composed of a coarse root filter and several higher resolution part filters as shown in Fig. 7, and adopts the deformation cost provided by the DPM detector to restrict the displacement of kernels during human tracking. Moreover, the proposed algorithm iteratively shifts the

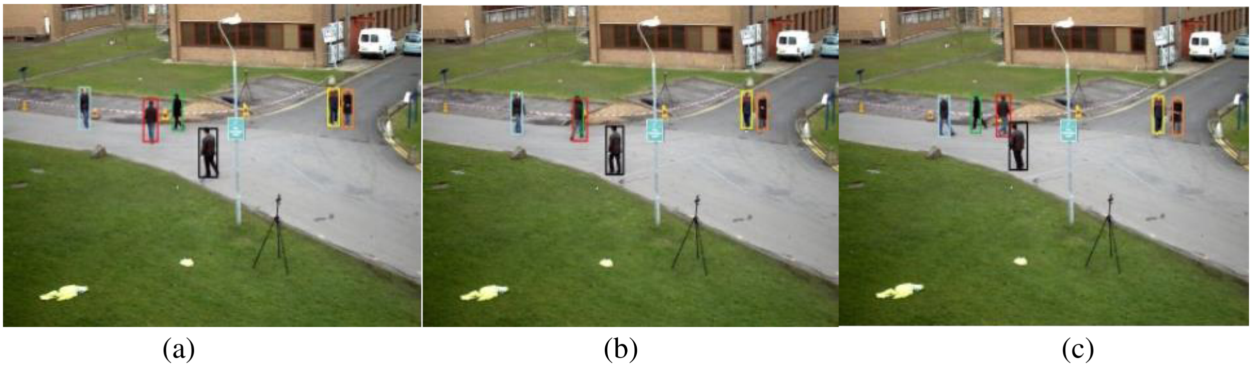


Fig. 6 Illustration of tracking multiple human targets in an outdoor scene. **a** Human targets tracked with different color *rectangle* bounding boxes. **b** Two human targets tracked successfully under fully mutual occlusion with a *red/green rectangle* bounding box. **c** Two human targets split correctly after fully mutual occlusion with a *red/green rectangle* bounding box[11]

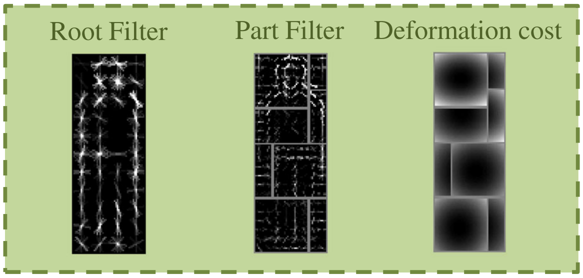


Fig. 7 Illustration of a pre-trained DPM. (*Left*) A coarse DPM root filter. (*Middle*) Several higher-resolution DPM part filters. (*Right*) Deformation cost associated with each DPM part filter [30]

kernels based on kernel histogram (i.e., spatially weighted color histogram) and histogram of oriented gradient (HOG) [31] in each video frame, and hence enables a robust and efficient human tracking solution without training required. In short, KT can achieve effective and robust as well as real-time human tracking by selecting excellent kernel function and sufficient human object representation. However, when a pedestrian move too fast or is totally occluded for a long time, the KT tends to lose the tracked human target.

3.2 Discriminative trackers

Discriminative trackers are another widely applied human tracking technique within a camera. Based on different disposal method of the human objects' association, the joint multiple-human tracking can be divided into three categories, i.e., JPDAF, MHT, and FNF. Table 3 lists qualitative comparison of discriminative trackers-based human tracking within a camera.

3.2.1 JPDAF

JPDAF is one of widely used techniques for data association in multi-target tracking. It jointly achieves multi-target tracking by associating all measurements with each track, where a track is defined through a sequence of measurements assumed to derive from the same object. Occlusion between tracked objects is one of the most difficult problems in multi-target tracking. To solve the issue, Rasmussen et al. [18] propose to track complex visual objects based on the JPDAF algorithm, where a related technique called Joint Likelihood Filter (JLF), i.e., relating the exclusion principle at the heart of the JPDAF to the method of masking out image data, is used to deal with occlusions between tracked objects. However, this method calls for very high computational requirements with the number of associated objects increasing. To take full advantage of more available information to further improve the tracking performance, Schulz et al. [19] propose sample-based JPDAF for tracking multiple moving human objects using a mobile robot, where the JPDAF algorithm is directly applied to the sample sets of the individual particle filter to determine the correspondence between the individual object and measurement. Moreover, the proposed approach adopts different features extracted from consecutive sensor measurements to explicitly deal with occlusions. However, the proposed method adopts fixed sample sizes for the particle filters, and randomly introduces samples whenever a new human object has been discovered. Therefore, more intelligent sampling techniques may result in improved results and faster convergence. To better deal with complex inter-target occlusion problems, with the aid of clustering process and extracted image features, Naqvi et al. [20] propose clustering and JPDAF for coping with occlusions in multi-target

Table 3 Qualitative comparison of discriminative trackers-based human tracking within a camera

Item No.	Used discriminative trackers	Speed	Occlusion	Scale change	Shape deformation
1	JLF-based JPDAF (Rasmussen et al. [18])	Low	√	√	√
2	Sample-based JPDAF (Schulz et al. [19])	Low	√	√	√
3	Clustering-based JPDAF (Naqvi et al. [20])	Low	√	√	√
4	JPDAF revisited (Rezatofighi et al. [21])	Moderate	√	√	√
5	Reliability measure-driven MHT (Zúñiga et al. [22])	High	√	√	√
6	MHT revisited (Kim et al. [23])	Moderate	√	√	√
7	Multiple association-based MHT (Joo et al. [24])	High	√	√	√
8	Hierarchical MHT (Zulkifley et al. [25])	Low	√	√	√
9	EOM-based FNF (Zhang et al. [26])	High	√	√	√
10	Greedy algorithms-based FNF (Pirsiavash et al. [27])	High	√	√	√
11	Lagrangian relaxation-based FNF (Butt et al. [28])	High	√	√	√
12	Multi-way data association-based FNF (Wu et al. [29])	Low	√	√	√

Symbol √ means that the used discriminative trackers-based human tracking within a camera can deal with the situations of occlusion, scale change, and shape deformation tracking. More specifically, the proposed algorithm adopts the variational Bayesian method for grouping measurements into clusters, and then uses a JPDAF technique to associate measurements to targets based on clustering image features; occlusion problems can thus be dealt with more effectively in multi-target tracking. However, this method is difficult to deal with numerous targets and measurements such as multiple human objects tracking in crowded scenes. To overcome this drawback, Rezatofighi et al. [21] revisit the JPDAF technique and propose a novel solution in formulating the problem as an integer linear program, which is embedded in a simple tracking framework. More specifically, the proposed method reformulates the calculation of individual JPDA assignment scores as a series of integer linear programs, and approximates the joint score by the m-best solutions, which is efficiently calculated by using a binary tree partition method, and hence addresses the issue of high computational complexity associated with JPDAF without forfeiting tracking performance. Such an example of tracking multiple human targets in a crowded scene is shown in Fig. 8. In short, the JPDAF is a good technique for data association in multi-target tracking, but it is very difficult to effectively track variable number of objects, such as a new object entering the field of view (FOV) or a tracked object exiting the FOV. Also, the JPDAF establishes the targets' correspondence using only two frames information; sometimes it will inevitably bring an incorrect correspondence.

3.2.2 MHT

MHT is another widely used technique for data association in multi-target tracking. It maintains several correspondence hypotheses for each object at each video frame and establishes the targets' correspondence through several frames of observations. However, the MHT has very high computational load since it exhaustively enumerates all possible associations. To reduce the computational requirement, Zúñiga et al. [22] propose a real-time MHT-based multi-human tracking approach, which can reliably track multiple human objects even in noisy environments. The proposed approach takes advantage of a dual object model through combining 2D with 3D features through reliability measures to generate tracking hypotheses of the moving human objects in the scene. Moreover, the proposed approach can manage many-to-many human objects' correspondences in real time. Kim et al. [23] revisit the MHT technique in a tracking-by-detection framework and propose a novel and more efficient MHT algorithm, which embeds online learned appearance models for each track hypothesis through a regularized least squares framework, and hence achieves pruning the hypothesis space more effectively and accurately so as to reduce the ambiguities in data association. However, the above MHT algorithm is still difficult to deal with complex interactions between the objects. To handle the issue, Joo et al. [24] propose multiple association-based MHT algorithms, relaxing the association constraint of conventional MHT to allow association of a single target with multiple measurements and multiple targets with a single measurement. More specifically, the proposed method regards the data association among multiple objects as a minimum weight bipartite graph edge, which is defined as a subset of edges such that each vertex is incident on at least one edge and the sum of the weights in the subset of edges is minimum, given an edge weighted graph. In addition, they develop a polynomial-time algorithm to generate only the best multiple association hypotheses, achieving robust and real-time target tracking. Zulkifley et al. [25] propose hierarchical two-level MHT for multiple-object tracking. The first level adopts foreground segmentation detection and clusters optical flow detection to generate observations so as to obtain stable velocity values and to filter out false track. The second level combines the outputs of the first-level with two additional virtual measurements based on appearance modeling and a big foreground blob to find the best combination of the observations. In short, the MHT algorithm has a wider practical application in multi-target tracking; it not only can track variable number of objects, but also can deal with the occlusion problem.

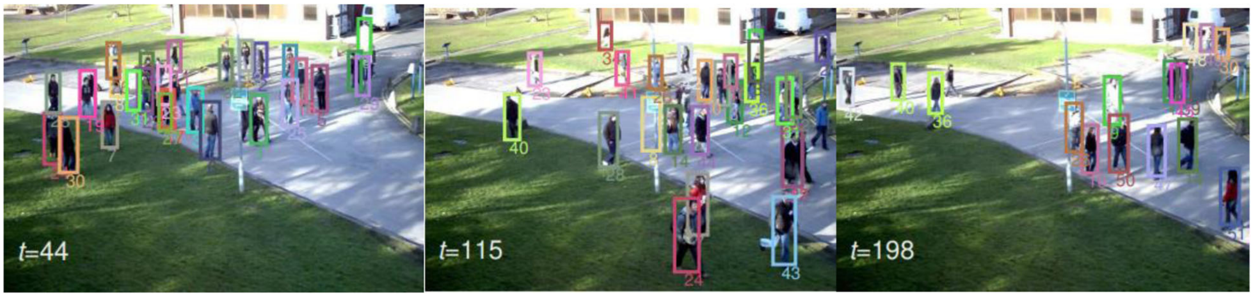


Fig. 8 Illustration of tracking multiple human targets in a crowded scene [21]

However, it has vitally high computational requirement, especially with the increased number of associated objects.

3.2.3 FNF

It becomes more and more popular in recent years to solve target association problems based on FNF, which is widely applied to multiple target tracking. Zhang et al. [26] propose an explicit occlusion model (EOM)-based minimal cost FNF to achieve robust multi-human tracking. The proposed approach maps the maximum a posteriori (MAP) data association problem into a cost-flow network with a non-overlap constraint on trajectories and adopts a min-cost flow algorithm to find the global optimal trajectory association in the network, given a set of human object detection results in each video frame as input observations, where observation likelihood and transition probabilities are modeled as flow costs, and non-overlapping trajectory hypotheses are modeled as disjoint flow paths. In addition, the proposed approach constructs an EOM through adding occlusion nodes and constraints to the network to solve long-term inter-object occlusion problems, and thus achieves real-time and robust multi-human tracking. Following the min-cost flow approach of EOM, Pirsiavash et al. [27] use a cost function that needs estimating the number of tracks, the objects' birth (i.e., a new object entering the FOV) and death state (i.e., a tracked object exiting the FOV) to formulate the computational problem of multi-object tracking. A greedy but globally optimal algorithm, which adopts shortest path computations based on a min-cost flow framework, is used for tracking a variable number of human objects. Such an example of tracking variable number of human objects in an outdoor scene is shown in Fig. 9. However, the above methods do not allow for path smoothness constraints. To solve the issue further, Butt et al. [28] develop a graph formulation that allows for encoding constant velocity constraints to evaluate the path smoothness over three adjacent frames, where candidate match pairs of observations are viewed as nodes in the graph, allowing each graph edge to encode an observation-based cost, and adopt the principle of Lagrangian relaxation to form a modified-cost network framework for global multi-human tracking. However, the above methods impose a constraint that one measurement is associated with only one target, i.e., one-to-one data association. To deal with many-to-one or one-to-many data associations, Park et al. [29] propose a general formulation called binary integer programming to handle a min-cost data association problem among target-measurement data associations through one-to-one, many-to-one, and one-to-many data associations (also called multi-way data associations) to track multiple interacting targets in video frames. The proposed method adopts Lagrangian dual relaxation to solve the binary integer programming problem, and hence achieves integer-valued solution with smaller duality gap than classical linear programming (LP) relaxation so as to improve the accuracy of data associations. However, the multi-way data associations are difficult to achieve real-time multiple human tracking. In short, the FNF-based tracking performance highly depends on the reliable detection. When the missing detection or long-time occlusion occurs, the tracking performance deteriorates significantly.

4 Human tracking across non-overlapping cameras

Human tracking across non-overlapping cameras establishes detected/tracked human objects' correspondence between two non-overlapping cameras so as to successfully perform label handoff. Based on the approaches used for target matching, human tracking across cameras can be divided into three main categories, human re-id, CLM-based tracking, and GM-based tracking.

For human re-id, which is to identify whether a human taken from one camera is the same as one taken from another camera or not. Human image-pair captured in two different cameras often varies greatly in appearance due to changes in illumination, viewpoint as well as intra-class variability in shape and pose. Such examples in VIPeR dataset [32] are shown in Fig. 10. The current research on the human re-id is primarily focused on two aspects [33]: one is extracting discriminative visual features to characterize human appearance and shape, the other is identifying suitable distance metrics that

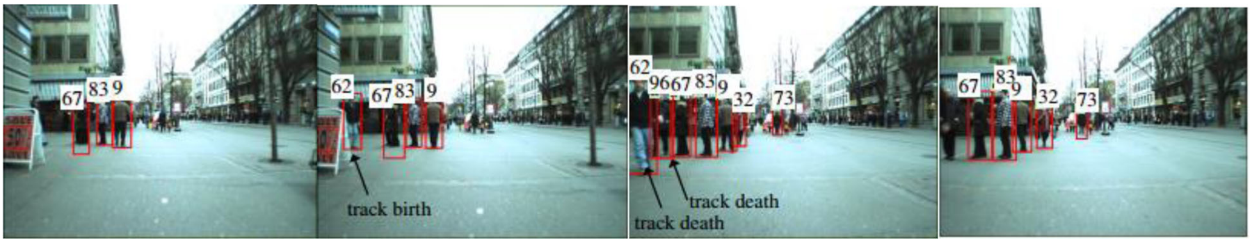


Fig. 9 Illustration of tracking variable number of human objects in an outdoor scene, including estimated track births and deaths [27]

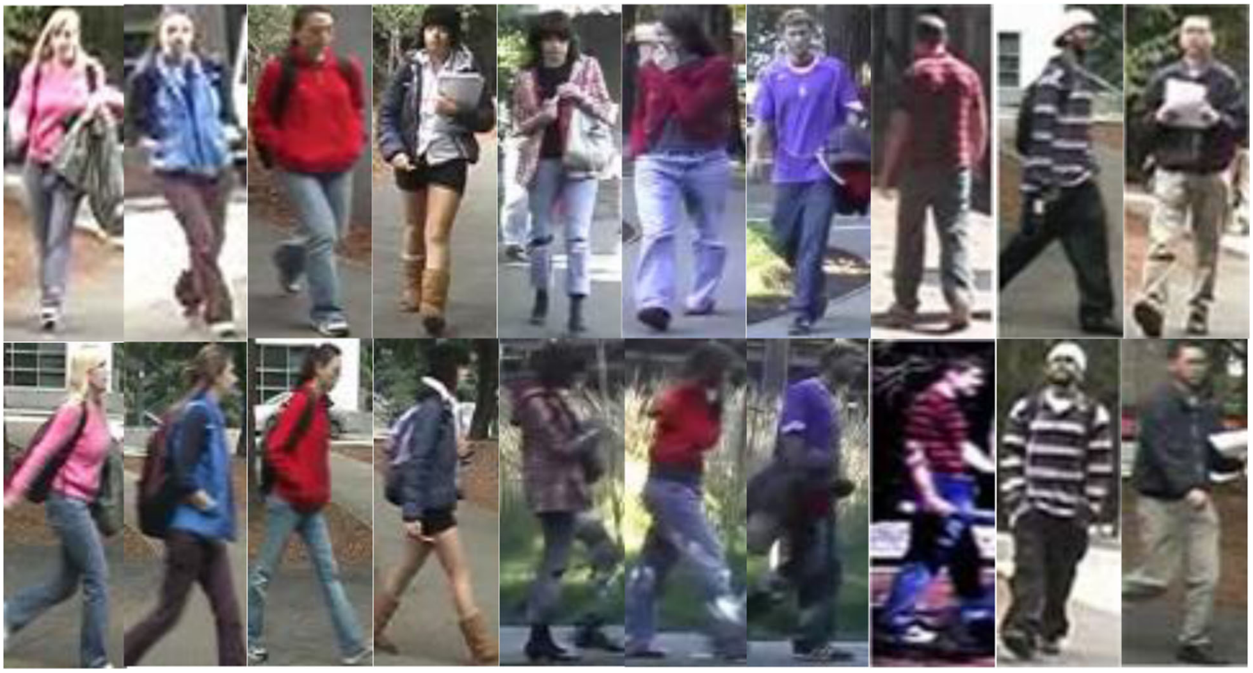


Fig. 10 Some human image-pair captured in two different cameras

maximize the likelihood of a correct correspondence. However, most visual features are either insufficiently discriminative for cross-view matching or insufficiently robust to viewpoint changes, resulting in a significant challenge for automated human re-id. Distance metric learning shifts the focus from capturing feature descriptors to learning distance metrics that maximize the human matching accuracy to improve human re-id performance. However, most distance metric learning requires pairwise supervised labeling of training datasets. It will become infeasible since the labeling needs a large amount of manual effort with the increased size of datasets or number of camera pairs.

For the CLM-based tracking, which is to track humans through establishing the link (correlation) models between two adjacent or among multiple neighboring cameras to compensate for the feature difference derived from different cameras. It is mainly applicable for tracking humans across multiple static cameras. The current research on the CLM-based tracking is primarily based on temporal and spatial relationships to reduce mismatch across cameras tracking, as well as appearance relationship to compensate for the appearance difference between two adjacent cameras. The CLM can be estimated in a supervised learning manner, i.e., with manually labeling the human objects' correspondence from given training data in advance; or an unsupervised learning manner, i.e., without manually labeling the human objects' correspondence from given training data. As a result, compared to the supervised learning-based CLM, which needs a lot of human labeling efforts, especially with the increased size of datasets or number of camera pairs, the unsupervised learning-based CLM is more feasible to achieve self-organized and scalable large-scale camera networks.

For the GM-based tracking, which is to track humans through a graph modeling technique to form a solvable GM based on input observations (detections, tracklets, trajectories or pairs) to deal with data association across cameras, where the GM is composed of nodes, edges, and weights and solved using an optimization solution through MAP estimation framework, to obtain optimal or suboptimal solutions. This tracking method can effectively track humans in complex scenes, such as occlusion, crowd, and interference of human appearance similarity. However, it is difficult to get the optimal solution of data association across cameras. Table 4 shows the list of the human tracking algorithms across non-overlapping cameras.

4.1 Human re-id

Human re-id is widely applied to human tracking across non-overlapping cameras. The current research on human re-id techniques mainly includes two aspects, i.e., feature extraction and distance metric learning. Table 5 lists quantitative comparison of human re-id across cameras on a quite challenging VIPeR benchmark dataset, using cumulative matching scores to evaluate the performance of human re-id, where the higher the cumulative matching scores are, the better the performance of human re-id is.

Table 4 Human tracking algorithms across non-overlapping cameras

Method	Description	Typical technique	Pros	Cons
Human re-id	To identify whether a human taken from one camera is the same as one taken from another camera or not	Feature extraction	Extracting discriminative and robust visual features help to improve human re-id accuracy	Difficult to find suitable feature combination to effectively describe human appearance
		Distance metric learning	Learning a distance metric helps to mitigate cross-view human appearances' variations.	Require manually pairwise labeling of training data
CLM-based tracking	To track humans through establishing the link (correlation) models between two adjacent or among multiple neighboring cameras	Supervised learning-based CLM	Easy to establish and learn CLM	Unfeasible to scale up to large-scale camera networks due to a mass of manually labeled efforts
		Unsupervised learning-based CLM	Help to achieve self-organized and scalable large-scale camera networks due to no need of human labeling efforts	Estimated CLM may decrease the accuracy due to higher outlier percentage.
GM-based tracking	To track humans through partite graph matching based on input observations (detections, tracklets, trajectories, or pairs)	MAP optimization solution framework	Human tracking in complex scenes such as occlusion, crowd, and interference of appearance similarity	It is difficult to get the optimal solution.

4.1.1 Feature extraction

Extracting discriminative and robust features from raw pixel data in an image/video has become one of the important tasks in human re-id. There are a lot of feature types proposed for human re-id, such as color [34], texture [35], shape [36], global features [34, 36], regional features [37], patch-based features [35], and semantic features [38]. In general, compared to other features, color feature is dominant under slight lighting changes since it is robust to changes in viewpoint. Texture or shape feature is stable under significant lighting changes, but they are subject to changes in viewpoint and occlusion. Global features, which reflect the global statistical characteristics of human appearance, have some invariance to changes in viewpoint and pose, but their discriminative power is not enough due to loss of spatial information which represents human object structure. Regional features and patch-based features increase the discriminative power further by taking into account the spatial information derived from partitioning the whole human region into several different regions, such as horizontal stripes, localized patches, and etc. Semantic features have better discriminative power and robustness to the cross-view variations. However, the semantic features require more labeling efforts, therefore, their generalization capability is limited. When executing cross-view human matching, the humans' appearance normally changes significantly due to the changes in illumination and viewpoint, therefore the use of a single feature to identify cross-view human objects is not enough. Most human re-id approaches benefit from integrating several features types to improve the cross-view human matching accuracy and robustness by taking advantage of the complementary nature among various features. Gray and Tao [39] propose the ensemble of localized features (ELF) to deal with viewpoint variations across cameras. More specifically, the ELF integrates RGB, YCbCr, HSV color features, and two kinds of texture features extracted through Schmid and Gabor filters with different radiuses and scales. An effective feature selection is performed through the AdaBoost machine learning algorithm to find the most discriminating features out of a large pool of color and texture features. Farenzena et al. [40] propose the Symmetry-Driven Accumulation of Local Features (SDALF) to describe human appearance across cameras. The SDALF encodes three complementary visual characteristics of the human appearance including the overall chromatic content represented through HSV color histogram, the spatial arrangement of colors into stable regions represented

Table 5 Quantitative comparison of human re-id across cameras on a quite challenging VIPeR benchmark dataset

Item no.	Used human re-id method	Rank on VIPeR			Reference
		1	10	20	
1	ELF	12.00	44.00	61.00	2008 ECCV [39]
2	SDALF	19.87	49.37	65.73	2010 CVPR [40]
3	ColorInv	24.21	57.09	69.65	2013 TPAMI [41]
4	SCNCD	37.80	81.20	90.40	2014 ECCV [42]
5	LOMO + XQDA	40.00	80.51	91.08	2015 CVPR [43]
6	FFN	51.1	91.4	96.9	2016 WACV [44]
7	KISSME	19.6	62.2	77.0	2012CVPR [45]
8	LFDA	24.18	67.12	82.00	2013 CVPR [46]
9	KLFDA	32.3	79.7	90.9	2014 ECCV [47]
10	MetricEnsb	45.9	88.9	95.8	2015 CVPR [48]
11	LSSL	47.8	87.6	94.2	2016 AAAI [49]
12	SCSP	53.5	91.5	96.6	2016 CVPR [50]

The cumulative matching scores (%) at rank 1, 10, and 20 are listed through maximally stable color regions (MSCR), and the presence of recurrent local motifs with high entropy represented through recurrent highly structured patches (RHSP), where the symmetry and asymmetry property is considered to handle viewpoint variations. Kviatkovsky et al. [41] propose to use color invariants (ColorInv) to perform human re-id. The ColorInv combines three component signatures over log color space including color histogram, covariance descriptor, and parts-based shape context (PartsSC), to describe human appearance, where the PartsSC, as an invariant shape descriptor using different parts of a human object, is used to describe the discriminative intra-distribution structure of color distributions. Yang et al. [42] propose salient color names-based color descriptor (SCNCD) for human re-id to deal with illumination changes across cameras, where the SCNCD and color histograms computed in four different color spaces, i.e., original RGB, rgb, $l_1l_2l_3$, and HSV, are fused to describe color features of human appearance. Note that the salient color names indicate that a color only has a certain probability of being assigned to several nearest color names, and that the closer the color name is to the color, the higher probability the color has of being assigned to this color name. Liao et al. [43] propose an effective feature representation of human appearance called Local Maximal Occurrence (LOMO) for human re-id, where the LOMO analyzes local color and texture features' horizontal occurrence and maximizes the occurrence so as to obtain a robust feature representation against viewpoint changes, based on HSV color histogram and scale invariant local ternary pattern (SILTP) texture descriptor. Such an illustration of the LOMO feature extraction method is shown in Fig. 11. Wu et al. [44] propose Feature Fusion Net (FFN) to describe human appearance for human re-id, where the FFN combines convolutional neural network (CNN) deep feature with handcrafted features, including color histogram computed in five different color spaces, i.e., RGB, HSV, YCbCr, Lab and YIQ, and Gabor texture descriptors with multi-scale and multi-orientation. The CNN deep feature is constrained by the handcrafted features through backpropagation to form a more discriminative feature fusion deep neural network. In short, discriminant multi-feature extraction with complementary nature helps to improve the accuracy of human re-id. However, the constructed feature vectors have very high dimension, resulting in very high computation requirement.

4.1.2 Distance metric learning

Since standard metrics, such as Euclidean distance for cross-view human matching in human re-id, based on the extracted features discussed previously, normally produce poor performance due to the potentially enormous changes in illumination, pose, and viewpoint. In order to mitigate cross-view variations and better identify more humans in human re-id, recent approaches [43, 45–50] are focused on learning an optimal metric model that aims to making features associated with the same human to be closer than features associated with different human objects. It is essential to learn a linear transformation that achieves a mapping from the original feature space to a new feature space so as to effectively execute human re-id. Mahalanobis metric learning is widely used to globally find the linear transformation of the feature space. Motivated by a statistical inference

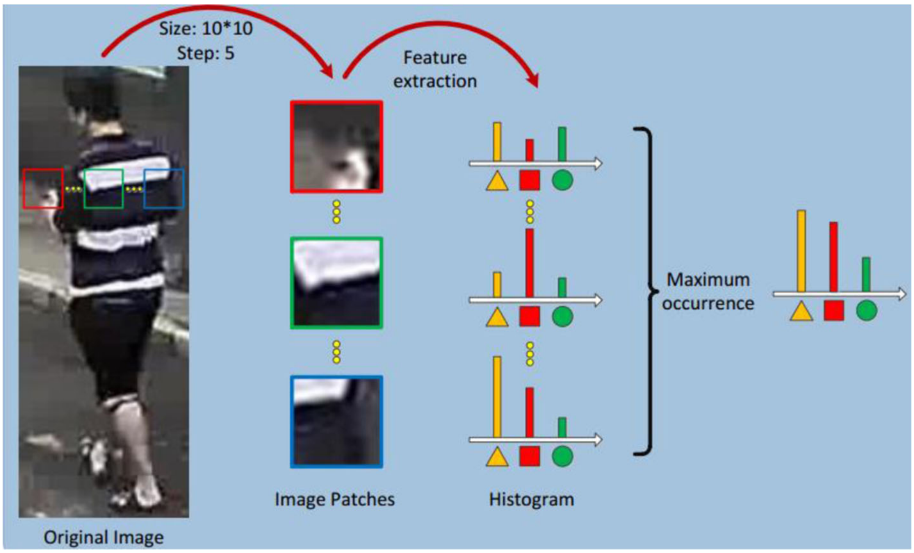


Fig. 11 Illustration of the LOMO feature extraction method [43]

perspective based on a likelihood-ratio test, Koestinger et al. [45] adopt equivalence constraints to learn a metric model called KISSME (keep it simple and straightforward metric). The proposed method only needs to compute two small-sized covariance matrices of dissimilar pairs and similar pairs, and thus is scalable to large datasets. Pedagadi et al. [46] adopt a low manifold distance metric learning framework through unsupervised PCA dimensionality reduction and supervised local fisher discriminant analysis (LFDA) dimensionality reduction, where the LFDA preserves the local neighborhood structure when maximizing between-class separation so as to achieve multi-class modality of the sample data, and the LFDA transformation is estimated via generalized eigenvalues. However, when this metric framework is applied to relatively small datasets, it may produce an undesirable compression of the most discriminative features. To solve this problem, by taking the merits from both kernel method and LFDA, Xiong et al. [47] further adopt kernel LFDA (KLFDA) to learn a metric model, where the KLFDA is a closed-form non-linear method that uses a kernel trick to handle large-dimensional feature vectors while maximizing a Fischer optimization criteria. The proposed method preserves discriminant features while achieving a better dimensionality reduction and takes full advantage of the flexibility in choosing the kernel to improve the accuracy of human re-id. However, its computational speed is relatively slow, especially when using non-linear kernel. Liao et al. [43] propose to learn a discriminant metric called cross-view quadratic discriminant analysis (XQDA), which aims to learn a low-dimensional subspace with cross-view data, and meanwhile learns a distance function in the low-dimensional subspace so as to measure the cross-view similarity. The proposed XQDA can be formulated as a generalized Rayleigh quotient, which can be solved by the generalized eigenvalue decomposition. However, the above proposed metric learning methods only adopt single metric learning model; integrating multiple metric learning models are thus also proposed in order to further improve the accuracy of human re-id. Paisitkriangkrai et al. [48] propose to learn to rank in human re-id with metric ensembles. More specifically, the proposed method first adopts several different features to train individual base metric of each feature using a linear KISSME and a non-linear KLFDA and then adopts two optimization approaches, i.e., relative distance-based approach and top recognition at rank-k, to learn weights of the base metrics. The two optimization approaches directly optimize a cumulative matching characteristic (CMC) curve, which is an evaluation measure commonly used in person re-id. The relative distance-based approach uses triplet information to optimize the relative distance, while the top recognition at rank-k approach maximizes the average rank-k recognition rate. Yang et al. [49] propose large-scale similarity learning (LSSL) using similar pairs for human re-id. More specifically, the proposed method jointly learns a Mahalanobis metric and a bilinear similarity metric using difference and commonness of an image pair to increase discrimination. Under a pair-constrained Gaussian assumption, the Gaussian priors (i.e., corresponding covariance matrices) of dissimilar pairs are obtained from those of similar pairs, and the application of a log likelihood ratio makes the whole learning process simple and fast and thus scalable to large datasets. However, the above metric learning methods just focus on a holistic metric, which discard the geometric structure of human objects and thus affect the discriminative power. To deal with the issue effectively, considering a relatively stable space distribution of human body parts such as head, torso, and legs, Chen et al. [50] propose spatially constrained similarity learning using polynomial feature map (SCSP) for human re-id. The proposed method, which combines a global similarity metric for the whole human body image region and multiple local similarity metrics for associating local human body parts regions using multiple visual cues, executes human matching across cameras based on multiple polynomial-kernel feature maps to represent human image pairs, which aims to learn a similarity function that could yield high score so as to measure the similarity between human image descriptors across cameras. Such an illustration of the similarity learning using spatial constraints based on polynomial-kernel feature map is shown in Fig. 12. In short, distance metric learning can improve the accuracy of human re-id effectively. However, most existing distance metric learning methods for human re-id follow a supervised learning framework, where a large number of labeled matching pairs are used for training, and hence severely limit the scalability in real-world applications. Moreover, the pre-trained distance metric model may not have better generalization ability.

4.2 CLM-based tracking

Since the human appearance may vary dramatically due to different viewpoints, poses, and illuminations, based on whether to use manually labeling the training data representing human correspondences or not, the research on the CLM-based tracking can be divided into two categories: the supervised learning-based CLM and the unsupervised learning-based CLM. Since most CLM-based tracking methods adopt different multiple camera tracking datasets, which is difficult to list all quantitative comparison of each CLM-based tracking method. Table 6 lists several quantitative comparison results of CLM-based tracking across non-overlapping cameras on NLPR datasets, using multiple camera

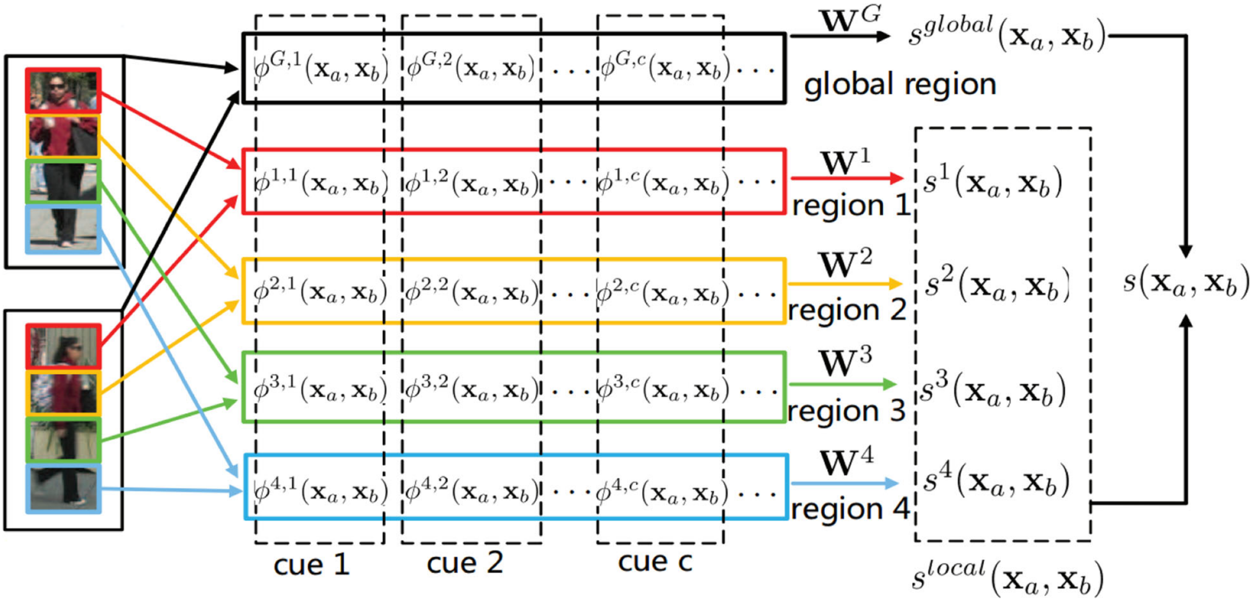

$\mathbf{W}^G$
$s^{global}(\mathbf{x}_a,\mathbf{x}_b)$
global region
$\phi^{G,1}(\mathbf{x}_a,\mathbf{x}_b)$
$\phi^{G,2}(\mathbf{x}_a,\mathbf{x}_b)$
$\phi^{G,c}(\mathbf{x}_a,\mathbf{x}_b)$
$\mathbf{W}^1$
region 1
$s^1(\mathbf{x}_a,\mathbf{x}_b)$
$\mathbf{W}^2$
region 2
$s^2(\mathbf{x}_a,\mathbf{x}_b)$
$\mathbf{W}^3$
region 3
$s^3(\mathbf{x}_a,\mathbf{x}_b)$
$\mathbf{W}^4$
region 4
$s^4(\mathbf{x}_a,\mathbf{x}_b)$
$s(\mathbf{x}_a,\mathbf{x}_b)$
cue 1
cue 2
cue c
$s^{local}(\mathbf{x}_a,\mathbf{x}_b)$

Fig. 12 Illustration of the similarity learning using spatial constraints based on polynomial-kernel feature map [50]

tracking accuracy (MCTA) to evaluate the performance of CLM-based tracking, where the higher the MCTA is, the better the performance of CLM-based tracking.

4.2.1 Supervised learning-based CLM

A supervised learning-based CLM, that is, the correspondences of pairs of individuals across every adjacent camera-pair are known in advance based on manually labeled training data, which can then be used to train a CLM. A number of studies have been reported to estimate the brightness transfer function (BTF), which is applied to compensating for the color difference between two adjacent cameras before computing the color feature distance between two observations. Javed et al. [51] propose to learn a low-dimensional subspace of the color brightness transfer function (BTF) from the training data for each camera-pairs using probabilistic PCA. However, this method depends on training data with a wide range of brightness values so as to accurately model the BTF, and it is difficult to meet this condition in a real-world scenario. To solve this problem, Prosser et al. [52] propose to adopt a cumulative brightness transfer function (CBTF) for mapping color information between adjacent cameras, which makes the best of the available color information from a very sparse training data set. This method can preserve uncommon brightness values in the training, resulting in more accurate representation of a color mapping function, therefore can help to improve the accuracy of human tracking across cameras. However, it only takes into account the color information and discards the spatial structural information for human representation. To cope with this problem, built upon the research of CRIPAC-MCT [51], Javed et al. [53] further adopt kernel density estimator to estimate the inter-camera space-time probabilities through computing the (e.g., walking) transition time values between pairs of correct correspondences based on the difference between the entry and exit time stamps. However, fully supervised learning usually requires a mass of manually labeled training data, which limits the scalability to more realistic open-world applications. To cope with this problem, Kuo et al. [54] adopt multiple instances learning (MIL) to learn an appearance affinity model, which is then integrated with the spatial-temporal information to train an improved inter-camera track association framework to tackle the target

Table 6 MCTA quantitative comparison of CLM/GM-based tracking across non-overlapping cameras on the existing NLPR datasets

Item no.	Used MCT method	CLM-based tracking	GM-based tracking	NLPR 1	NLPR 2	NLPR 3	NLPR 4	Reference
1	Duke MTMC	√	×	0.7967	0.7336	0.6543	0.7616	2016 ECCV [64]
2	USC	√	×	0.9152	0.9132	0.5163	0.7052	2014 WACV [55]
3	SG-CRF	×	√	0.8383	0.8015	0.6645	0.7266	2016 TCSVT [61]
4	CRIPAC-MCT	×	√	0.6617	0.5907	0.7105	0.5703	2014 ICIP [62]
5	EG	×	√	0.8353	0.7034	0.7417	0.3845	2016 TCSVT [63]

Symbols √ and × mean whether CLM/GM based tracking is used or not handover tasks across cameras. In addition, people often walk in groups in crowded scenes, thus group information is also applied to appearance matching across cameras. Cai et al. [55] propose context information including spatio-temporal context and relative appearance context for non-overlapping inter-camera human tracking. The spatio-temporal context indicates a way of collecting samples for discriminative appearance learning, and the relative appearance context using RGB color histograms and histogram of gradients as appearance features models inter-object appearance similarities for people walking in proximity. The proposed method can distinguish visually very similar human targets and hence obviously improves human tracking accuracy across non-overlapping cameras. In short, the supervised learning-based CLM helps to achieve robust human tracking across non-overlapping cameras. However, it is unfeasible to scale up to large-scale camera networks due to a mass of manually labeled efforts.

4.2.2 Unsupervised learning-based CLM

Contrary to the supervised learning-based CLM, an unsupervised learning-based CLM, that is, the correspondences of pairs of individuals across every adjacent camera-pair are unknown in advance, which can still be estimated and then be used to train a CLM. The time-space and appearance relationships between adjacent cameras are usually used to learn the CLM across camera-pairs. Makris et al. [56] adopt the cross-correlation of the exit and entry time stamps of the training data to estimate the transition time distribution. However, they only consider the single-mode distribution, thus it is difficult to describe most cases in the real world. Gilbert et al. [57] propose an incremental learning method to model the color variations and the transition time distribution between cameras. The proposed method allows to increase human tracking accuracy over time without any supervised input. However, they consider all the possible correspondences within a given time window including the true and false correspondences, and hence large amount of noises are produced due to a large number of false correspondences during the whole estimation process, resulting in unreliable model estimation. Chu et al. [58] adopt transition time distribution and brightness transfer function, based on space–time relationship and holistic and regional color/texture information, respectively, between a pair of directly connected cameras, to estimate a CLM. A permutation matrix is introduced as an intermediate variable to be solved by using a deterministic annealing and the barrier method. This approach also takes into account the outliers, which refers to those people who depart from a camera without entering the other connected camera, or enter into a camera without coming from the other connected camera. In order to make the estimated CLM more accurately and adapt to environmental changes, by effective estimation of the feature fusion weights, the CLM can be persistently updated based on the human re-id results during tracking in the testing stage. The proposed CLM estimation method is applied in a deployed 4-camera real-world scenario with non-overlapping views, whose camera topology is shown in Fig. 13, achieving 79.5% tracking accuracy out of 20 min (more than 280 people) of video testing. However, their approach of coping with the outliers only considers a link of a pair of directly connected cameras. In many real-world camera networks, there are often several links due to multiple directly connected cameras; in this case, their estimated CLM will decrease the accuracy due to higher outlier percentage. In order to solve this problem, built upon the research of Ref. [58], Lee et al. [59] propose to combine multi-camera links and build bidirectional transition time distribution during the estimation of the CLM between directly connected camera pairs, and several camera link models are simultaneously estimated for the same deployed 4-camera real-world camera network with non-overlapping views in the presence of the outliers, resulting in more accurate camera link model and achieving 87.3% tracking accuracy. In short, the unsupervised learning-based CLM helps to achieve robust human tracking across non-overlapping cameras, and can be easily applied to real-world systems with continuous updates of the link models when the conditions between cameras change. Moreover, it is feasible to achieve self-organized and scalable large-scale camera networks due to no need of human labeling efforts.

4.3 GM-based tracking

GM-based tracking using the optimization framework is also applied to human tracking across non-overlapping cameras. Javed et al. [60] propose to establish human objects' correspondences across non-overlapping cameras through the MAP estimation framework based on human motion trends and appearance of human objects. More specifically, the proposed method adopts Parzen windows, i.e., kernel density estimators, to estimate inter-camera space-time probabilities from the training data between each pair of cameras, and models the changes in human appearance using the distances between color models. To estimate the human correspondences across non-overlapping cameras, the proposed method then models the issue of finding the hypothesis that maximizes the MAP as finding the path of a directed graph. In addition, to keep up with the changing human motion and appearance patterns, the proposed method continuously updates the learned parameters during the human tracking across non-overlapping cameras. However, the above method only focuses on

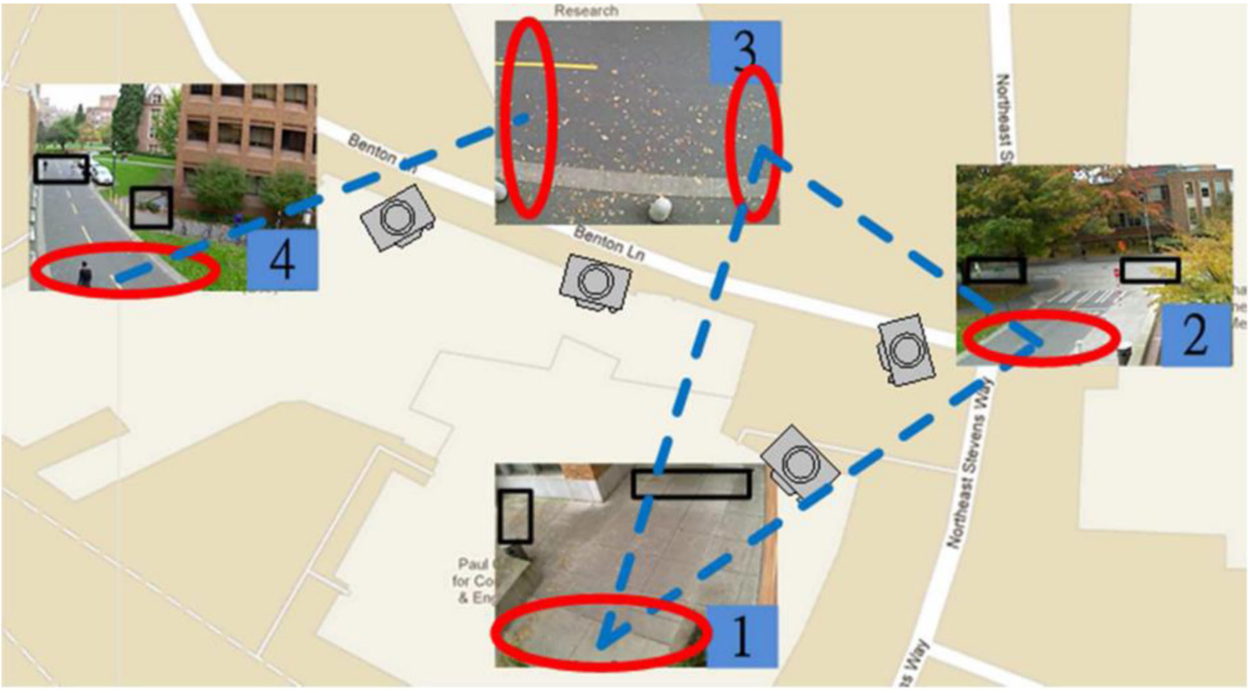


Fig. 13 Camera topology. *Blue broken lines* denote four links, and *red ellipses* denote the corresponding entry or exit zones. *Black rectangles* are the other entry or exit zones that have no any link between both cameras [58]

appearance and spatio-temporal cues, Chen et al. [61] combine high-level contextual information called social grouping behavior with traditionally used appearance and spatiotemporal cues into a non-overlapping inter-camera human tracking system, and adopt an online learned conditional random field model that minimizes a global energy cost to associate tracks from the same person of different cameras, and hence effectively achieve human tracking across non-overlapping cameras. The above proposed methods adopt the trajectories obtained from single camera human tracking to achieve inter-camera data association, and hence the overall tracking performance depends on the results of single camera human tracking, especially in challenging scene videos, the direct disturbance of false positives and fragments will seriously decrease the overall tracking performance. Such an example of human tracking across non-overlapping cameras on NLPR 4 is shown in Fig. 14. To deal with human tracklet mismatching and missing issues (as shown in Fig. 15) across non-overlapping cameras, Chen et al. [62] propose a global tracklet association for human tracking across non-overlapping cameras to improve the overall tracking performance. More specifically, the proposed method adopts fragmentary tracklets as the inputs based on a piecewise major color spectrum histogram representation (PMCSHR) and models a global tracklet association as a global MAP problem, which is mapped into a cost-flow network and solved by a min-cost flow algorithm. In addition, to better achieve tracklet matching across multiple camera views, the minimum uncertainty gap-based measurement, i.e., using the lowest and highest similarity to define the lower and upper bounds of the similarity for two tracklets to obtain a distance metric, is applied to computing the matching result of two tracklets' PMCSHRs. Built upon the research of PMCSHR [62], Chen et al. [63] equalize similarity metrics in the global graph based on appearance and motion features, and hence further reduce the number of mismatch errors in non-overlapping inter-camera human tracking so as to further improve human tracking performance across non-overlapping cameras. Table 6 lists several quantitative comparison results of GM-based tracking across non-overlapping cameras on NLPR datasets, using multiple camera tracking accuracy (MCTA) to evaluate the performance of GM-based tracking, where the higher the MCTA is, the better the performance of GM-based tracking.

5 Conclusions

This paper provides an extensive review of existing research efforts on human tracking over camera networks, covering all the core image/vision technologies, such as generative trackers, discriminative trackers, human re-id, CLM-based tracking, and GM-based tracking. We discuss the most recent development of these technologies

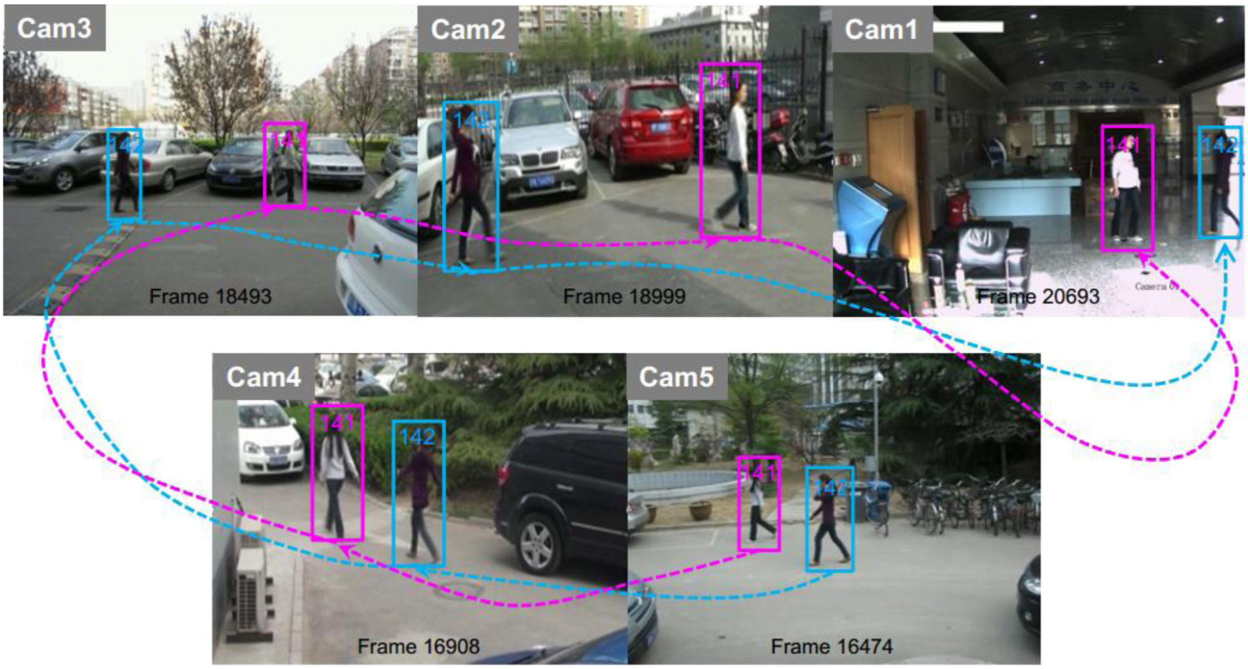
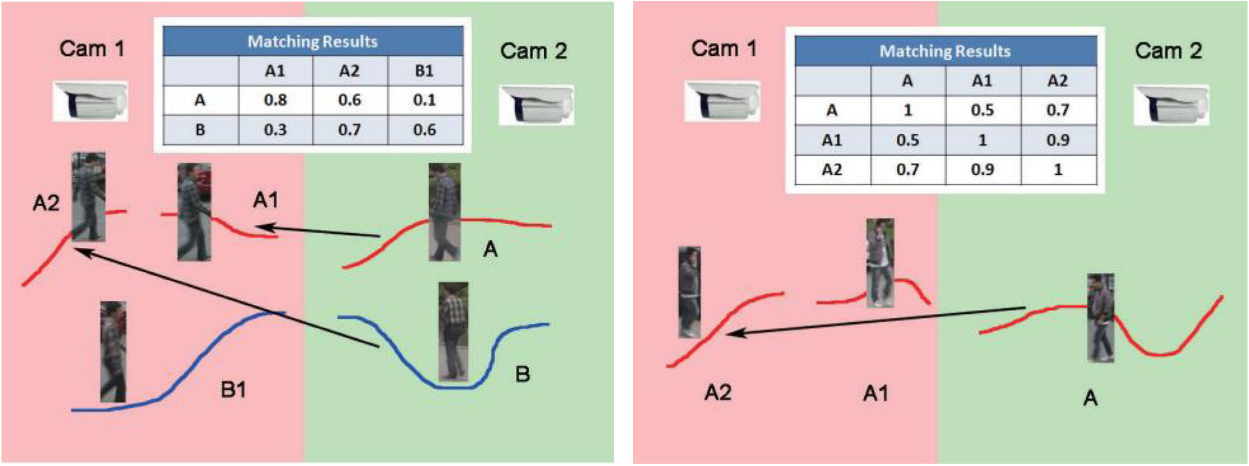


Fig. 14 Illustration of human tracking results across non-overlapping in cameras on NLPR 4. *Bounding boxes* with the same color indicate the same human, and the *dashed lines* illustrate the trajectories generated by human targets walking across different cameras [61]

and compare pros/cons of different solutions. In spite of the great progress made on the human tracking over camera networks including human tracking within a camera and human tracking across non-overlapping cameras, there are still many technical challenges that need to be resolved, especially for real-world camera networks. For example, (1) when a human target is totally occluded for a long time or the background is extremely complex in the same camera scene, it is difficult to extract robust and discriminant features that denote human targets, resulting in the decline of performance for human tracking within a camera; (2) extracting robust and discriminant features adaptive to changes in illumination, viewpoint, occlusion, background clutter, and image quality/resolution across non-overlapping cameras, is still a challenging issue; (3) most learned distance metric models from an initial annotated camera-pair in human re-id are difficult to expand or



Matching Results

	A1	A2	B1
A	0.8	0.6	0.1
B	0.3	0.7	0.6

Matching Results

	A	A1	A2
A	1	0.5	0.7
A1	0.5	1	0.9
A2	0.7	0.9	1

(a) Human tracklet mismatching

(b) Human tracklet missing

Fig. 15 Illustration for human tracklet mismatching and missing. *Blue* and *red* lines indicate two human targets, and *arrows* show the best matching. Human target B is mismatched to human tracklet A2 in (**a**). Human tracklet A1 is missing in (**b**) [62]

adapt to a new camera-pair due to differences in illumination and viewpoint. Moreover, these models cannot be updated adaptively with the real-world environment changes. Also, it is impractical to manually label a large number of training data from every camera-pairs for a large camera networks; (4) so far, the performance of human re-id is still far from satisfactory, for example, the rank-1 accuracy of state-of-the-art, based on cumulative matching scores evaluation, is less than 60% on the representative VIPeR dataset, which will bring huge challenges for the human tracking across non-overlapping cameras when spatio-temporal reasoning between cameras is unreliable, especially for the human tracking across multiple moving cameras due to the fact that the mapping between two cameras will change with the cameras' movement; (5) the larger the spatio-temporal separation between camera views is, the greater the chance that human may appear with more appearance changes in different camera views is, resulting in difficulty to track human across non-overlapping cameras; (6) most existing research efforts on human tracking across non-overlapping cameras are based on available small camera networks composed of no more than five cameras; how to expand these techniques for human tracking over larger-scale camera networks.

In terms of the above unsolved technical challenges of tracking human over camera networks, future research directions on human tracking over camera networks can be summarized as follows:

1) Robust and discriminant feature fusion adaptive to camera scene changes for human tracking over camera networks.
2) Robust and discriminant spatio-temporal and appearance context information for inter-camera human tracking.
3) Effective distance metric learning fusion to improve human re-id accuracy.
4) Online human tracking across non-overlapping cameras using unsupervised learning.
5) Effective global data association for human tracking over camera networks.
6) Human tracking on larger-scale camera networks as well as benchmark datasets and comprehensive experimental evaluations on larger-scale camera networks.

Funding

This work was supported by the National Natural Science Foundation of China (61373084), Anhui Key Research and Development Plan Project (1704e1002208), and Anhui Natural Science Research Project (KJHS2016B03).

Authors' contributions

LH conceived and designed the study, and wrote the manuscript. WGW and J-NH provided the technical advice. RM revised the manuscript. MYY and KH provided some references. All authors read and approved the final manuscript.

Authors' information

Li Hou received her BS degree in Communication Engineering and MS degree in Power Electronics from Liaoning University of Technology in 2003 and 2006, respectively. She joined the School of Information Engineering of Huangshan University in 2006. She is currently a PhD candidate at the School of Communication and Information Engineering of Shanghai University. Her current research interests include computer vision, machine learning, and video/image processing.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. [2]School of Information Engineering, Huangshan University, 245041 Huangshan, China. [3]Institute of Smart City, Shanghai University, 200444 Shanghai, China. [4]Department of Electrical Engineering, University of Washington, Seattle, WA 98195, USA.

Received: 6 October 2016 Accepted: 25 May 2017
Published online: 05 June 2017

References

1. A Yilmaz, O Javed, M Shah, Object tracking: a survey. ACM Comput. Surv. (CSUR) **38**(4), 1–13 (2006)
2. AW Smeulders, DM Chu, R Cucchiara, S Calderara, A Dehghan, M Shah, Visual tracking: an experimental survey. IEEE Trans. Pattern Anal. Mach. Intell. **36**(7), 1442–1468 (2014)
3. Y Wu, J Lim, MH Yang, "Online object tracking: a benchmark," IEEE Conf. Computer Vision and Pattern Recognition, Portland, USA, Jun. 2013.
4. DS Jang, HI Choi, Active models for tracking moving objects. Pattern Recogn. **33**(7), 1135–1146 (2000)
5. DS Jang, SW Jang, HI Choi, 2D human body tracking with structural kalman filter. Pattern Recogn. **35**(10), 2041–2049 (2002)
6. SK Weng, CM Kuo, SK Tu, Video object tracking using adaptive kalman filter. J. Vis. Commun. Image Represent. **17**(6), 1190–1208 (2006)
7. X Li, et al. A multiple object tracking method using Kalman filter, IEEE International Conf. Information and Automation, Harbin, China, Jun. 2010.
8. DN Cong, et al. Robust visual tracking via MCMC-based particle filtering, IEEE Int. Conf. Acoustics, Speech and Signal Processing, Tokyo, Japan, Mar. 2012.
9. X Zhang, W Hu, S Maybank, A smarter particle filter, Asian Conf. Computer Vision Computer, Xi'an, China, Sep. 2009.
10. K Meshgi et al., An occlusion-aware particle filter tracker to handle complex and persistent occlusions. Comput. Vis. Image Underst. **150**(9), 81–94 (2016)
11. B Yang, R Yang, Interactive particle filter with occlusion handling for multi-target tracking, IEEE Int. Conf. Fuzzy Systems and Knowledge Discovery, Zhangjiajie, China, Aug. 2015.
12. C Liu, C Hu, JK Aggarwal, "Eigenshape kernel based mean shift for human tracking," IEEE Int. Conf. Computer Vision Workshops, Barcelona, Spain, Nov. 2011.
13. CT Chu, JN Hwang, HI Pai, KM Lan, Tracking human under occlusion based on adaptive multiple kernels with projected gradients. IEEE Trans. Multimedia **15**(7), 1602–1615 (2013)
14. J Fang, J Yang, H Liu, Efficient and robust fragments-based multiple kernels tracking. Int. J. Electron. Commun. **65**(1), 915–923 (2011)
15. L Hou, et al., Deformable multiple-kernel based human tracking using a moving camera, IEEE Int. Conf. Acoustics, Speech and Signal Processing, Brisbane, Australia, Apr. 2015.
16. L Hou et al., Robust human tracking based on DPM constrained multiple-kernel from a moving camera. J. Signal Process. Syst. Signal ImageVideo Technol. **86**(1), 27–39 (2017)
17. M Paul, SM Haque, S Chakraborty, Human detection in surveillance videos and its applications-a review, EURASIP Journal on Advances in Signal Processing, no. 1, 2013: 176.

18. C Rasmussen, G Hager, Probabilistic data association methods for tracking complex visual objects. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 560–576 (2001)
19. D Schulz, W Burgard, D Fox, AB Cremers, People tracking with mobile robots using sample based joint probabilistic data association filters. Int. J. Robot. Res. **22**(2), 99–116.16 (2003)
20. SM Naqvi, L Mihaylovay, JA Chambers, Clustering and a joint probabilistic data association filter for dealing with occlusions in multi-target tracking, Int. Conf. Information Fusion, Istanbul, Turkey, Jul. 2013.
21. S Hamid Rezatofighi, et al., Joint probabilistic data association revisited, IEEE International Conf. Computer Vision, Santiago, Chile, Dec. 2015.
22. MD Zúñiga, F Brémond, M Thonnat, Real-time reliability measure-driven multi-hypothesis tracking using 2D and 3D features, EURASIP Journal on Advances in Signal Processing, no. 1, 2011, pp. 1–21.
23. C Kim, F Li, A Ciptadi, JM Rehg, Multiple hypothesis tracking revisited, IEEE Int. Conf. Computer Vision, Santiago, Chile, Dec. 2015.
24. SW Joo, R Chellappa, A multiple-hypothesis approach for multiobject visual tracking. IEEE Trans. Image Process. **16**(11), 2849–2854 (2007)
25. MA Zulkifley, B Moran, Robust hierarchical multiple hypothesis tracker for multiple-object tracking. Expert Syst. Appl. **39**(16), 12319–12331 (2012)
26. L Zhang, Y Li, and R Nevatia. Global data association for multi-object tracking using network flows, IEEE Conf. Computer Vision and Pattern Recognition, Anchorage, USA, Jun. 2008.
27. H Pirsiavash, D Ramanan, CC Fowlkes, Globally optimal greedy algorithms for tracking a variable number of objects, IEEE Conf. Computer Vision and Pattern Recognition, Providence, USA, Jun. 2011.
28. AA. Butt, RT Collins, Multi-target tracking by lagrangian relaxation to min-cost network flow, IEEE Conf. Computer Vision and Pattern Recognition, Portland, USA, Jun. 2013.
29. C Park et al., Minimum cost multi-way data association for optimizing multitarget tracking of interacting objects. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 611–624 (2015)
30. PF Felzenszwalb et al., Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
31. N Dalal, B Triggs, Histograms of oriented gradients for human detection, IEEE Conf. Computer Vision and Pattern Recognition, San Diego, CA, USA, Jun. 2005.
32. D Gray, S Brennan, H Tao, Evaluating appearance models for recognition, reacquisition, and tracking, IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), 2007.
33. X Wang, Intelligent multi-camera video surveillance: a review. Pattern Recogn. Lett. **34**(1), 3–19 (2013)
34. ED Cheng, M Piccardi, Matching of objects moving across disjoint cameras, IEEE Int. Conf. Image Processing, Atlanta, USA, Oct. 2006.
35. W Li, X Wang, Locally aligned feature transforms across views, IEEE Conf. Computer Vision and Pattern Recognition, Portland, USA, Jun. 2013.
36. X Wang, et al., Shape and appearance context modeling, IEEE Conf. Computer Vision, Rio de Janeiro, Brazil, Oct. 2007.
37. DS Cheng, et al., Custom pictorial structures for re-identification, British Machine Vision Conference, Dundee, UK, Sep. 2011.
38. R Layne, T Hospedales, S Gong, Person re-identification by attributes, British Machine Vision Conference, Surrey, UK, Sep. 2012.
39. D Gray, H Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, European conf. Computer Vision, Marseille, France, Oct. 2008.
40. M Farenzena, et al., Person re-identification by symmetry-driven accumulation of local features, IEEE Conf. Computer Vision and Pattern Recognition, San Francisco, CA, USA, Jun. 2010.
41. I Kviatkovsky, A Adam, E Rivlin, Color invariants for person reidentification. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1622–1634 (2013)
42. Y Yang, et al., Salient color names for person re-identification, European Conf. Computer Vision, Zurich, Switzerland, Sep. 2014.
43. S Liao, Y Hu, X Zhu, and SZ Li, Person re-identification by local maximal occurrence representation and metric learning, IEEE Conf. Computer Vision and Pattern Recognition, Boston, MA, USA, Jun. 2015.
44. S Wu, YC Chen, X Li, AC Wu, JJ You, WS Zheng, An enhanced deep feature representation for person re-identification, IEEE Winter Conf. Applications of Computer Vision, Lake Placid, NY, USA, Mar. 2016.
45. M Koestinger, et al., Large scale metric learning from equivalence constraints, IEEE Conf. Computer Vision and Pattern Recognition, Providence, USA, Jun. 2012.
46. S Pedagadi, et al., Local fisher discriminant analysis for pedestrian re-identification, IEEE Conf. Computer Vision and Pattern Recognition, Portland, USA, Jun. 2013.
47. F Xiong, M Gou, O Camps, M Sznaier, Person re-identification using kernel-based metric learning methods, European Conf. Computer Vision, Zurich, Switzerland, Sep. 2014.
48. Paisitkriangkrai, C Shen, A Hengel, Learning to rank in person re-identification with metric ensembles, IEEE Conf. Computer Vision and Pattern Recognition, Boston, MA, USA, Jun. 2015.
49. Y Yang, S Liao, Z Lei, SZ Li, Large scale similarity learning using similar pairs for person verification, AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, Feb. 2016.
50. D Chen, Z Yuan, B Chen, N Zheng, Similarity learning with spatial constraints for person re-identification, IEEE Conf. Computer Vision and Pattern Recognition, Las Vegas, Nevada, USA, Jun. 2016.
51. O Javed, K Shafique, M Shah, Appearance modeling for tracking in multiple non-overlapping cameras, IEEE Conf. Computer Vision and Pattern Recognition, San Diego, USA, Jun. 2005.
52. B Prosser, S Gong, T Xiang, Multi-camera matching using bi-directional cumulative brightness transfer function, British Machine Vision Conference, Leed, UK, Sep. 2008.
53. O Javed, K Shafique, Z Rasheed, M Shah, Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views. Comput. Vis. Image Underst. **109**(2), 146–162 (2008)
54. CH Kuo, C Huang, R Nevatia, Inter-camera association of multi-target tracks by on-line learned appearance affinity models, European Conf. Computer Vision, Heraklion, Greece, Sep. 2010.
55. Y Cai, G Medioni, Exploring context information for inter-camera multiple target tracking, IEEE Winter Conf. Applications of Computer Vision (WACV), Colorado, USA, Mar. 2014.
56. D Makris, T Ellis, J Black, Bridging the gaps between cameras, IEEE Conf. Computer Vision and Pattern Recognition, Washington, USA, Jul. 2004.
57. A Gilbert, R Bowden, Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity, European Conf. Computer Vision, Graz, Austria, May 2006.
58. CT Chu, JN Hwang, Fully unsupervised learning of camera link models for tracking humans across non-overlapping cameras. IEEE Trans. Circuits Syst. Video Technol. **24**(6), 979–994 (2014)
59. YG Lee, JN Hwang, Z Fang, Combined estimation of camera link models for human tracking across non-overlapping cameras, IEEE Int. Conf. Acoustics, Speech and Signal Processing, Brisbane, Australia, Apr. 2015.
60. O Javed, Z Rasheed, K Shafique, M Shah, Tracking across multiple cameras with disjoint views, IEEE International Conf. Computer Vision, Nice, France, Oct. 2003.
61. X Chen, B Bhanu, Integrating social grouping for multi-target tracking across cameras in a CRF model," IEEE Trans. Circuits and Systems for Video Technology. doi:10.1109/TCSVT.2016.2565978.
62. W Chen, L Cao, X Chen, K Huang, A novel solution for multi-camera object tracking, IEEE Int. Conf. Image Processing, Paris, France, Oct. 2014.
63. W Chen, L Cao, X Chen, K Huang, An equalised global graphical model-based approach for multi-camera object tracking, IEEE Trans. Circuits and Syst. Video Technol. doi:10.1109/TCSVT.2016.2589619.
64. E Ristani, F Solera, R Zou, R Cucchiara, C Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, European Conf. Computer Vision, 2016.