Pattern Recognition 44 (2011) 1121–1136

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Spatial–temporal consistent labeling of tracked pedestrians across non-overlapping camera views

Guoyun Lian [a], Jianhuang Lai [a,*], Wei-Shi Zheng [b,c]

[a] School of Information Science and Technology, Sun Yat-sen University, 510006, China
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK
[c] Guangdong Province Key Laboratory of Information Security, Sun Yat-sen University, China

ARTICLE INFO

Article history:
Received 23 March 2010
Received in revised form
21 September 2010
Accepted 16 November 2010

Keywords:
Consistent labeling
Non-overlapping camera views
Bayesian model
CMCSHR
Optimal graph matching (OGM)

ABSTRACT

Tracking people across multiple cameras with non-overlapping views is a challenging task, since their observations are separated in time and space and their appearances may vary significantly. This paper proposes a Bayesian model to solve the consistent labeling problem across multiple non-overlapping camera views. Significantly different from related approaches, our model assumes neither people are well segmented nor their trajectories across camera views are estimated. We formulate a spatial–temporal probabilistic model in the hypothesis space that consists the potentially matched objects between the exit field of view (FOV) of one camera and the entry FOV of another camera. A competitive major color spectrum histogram representation (CMCSHR) for appearance matching between two objects is also proposed. The proposed spatial–temporal and appearance models are unified by a maximum-a-posteriori (MAP) Bayesian model. Based on this Bayesian model, when a detected new object corresponds to a group hypothesis (more than one object), we further develop an online method for online correspondence update using optimal graph matching (OGM) algorithm. Experimental results on three different real scenarios validate the proposed Bayesian model approach and the CMCSHR method. The results also show that the proposed approach is able to address the occlusion problem/group problem, i.e. finding the corresponding individuals in another camera view for a group of people who walk together into the entry FOV of a camera.

1. Introduction

In visual surveillance, it is not possible to monitor a wide interesting area using a single camera, because the field of view (FOV) of one camera is finite and the scene structure limits the visible area in one camera view [1]. Therefore, surveillance system for wide areas has to be deployed in a network of cameras. One of the major tasks of the camera network is to track objects across multiple cameras.

Tracking across multi-camera with overlapping field of views has been discussed in [2–8] by mainly using geometric information such as planar ground homography and the axis of an object [2,5], handoff table consisting of point correspondence between two views [3], the limits of FOV of each camera as visible in the other cameras [4], or appearance information [7,8].

However, in practice, it could be usually hard to completely cover large areas with overlapping cameras views due to limited resources. Thus, matching of moving objects across disjoint cameras is now becoming more and more important for any surveillance system. The task is challenging, since no continuous information is provided in this case and in consequence the observations of the same object can be widely separated in time and space. To address these challenges, in this paper, we present a unified Bayesian framework that solves the consistent labeling problem of people across non-overlapping camera views.

1.1. Related work

Tracking objects across multiple cameras over disjoint views is generally a consistent labeling problem, and existing methods are mainly based on spatio-temporal cue, visual cue, or their combination.

1.1.1. Spatial–temporal cue

Makris et al. [9] and Wang et al. [10] explore the spatio-temporal relationship by modeling trajectories of objects based on the activities information. Their work holds an assumption that trajectories are likely to correspond to the same object if they belong to the same activity. Similar work were also reported by Rahimi and Darrell [11] and Kettnaker and Zabih [12]. Rahimi and Darrell utilized location and velocity of an object observed by non-overlapping camera views to recover the trajectory most

* Corresponding author. Tel.: +86 013168313819.
E-mail addresses: lgyun2005@gmail.com (G. Lian), stsljh@mail.sysu.edu.cn (J. Lai), wszheng@ieee.org (W.-S. Zheng).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.11.011

compatible with the object dynamics [11]. Kettnaker and Zabih [12] proposed to use the transition time of objects across cameras for tracking. A Bayesian model was used to reconstruct the paths of the objects across multiple cameras. Performances of these methods greatly depend on how well the objects actually follow the estimated trajectory across camera views. Moreover, these methods need observe object correspondences over a long period of time in order to train their models.

1.1.2. Visual cue

Without acquiring a complete trajectory, several other work aims to perform consistent labeling across disjoint camera views using visual cue [13,14,1,15,16]. In these methods, the appearance color information of people is used. Visual cue, especially color, is always largely affected by lighting variations. To alleviate this problem, an illumination-tolerant appearance representation based on online k-means color clustering algorithm is introduced in [13,14], which can alleviate the typical illumination variations. They tracked people based on a major color spectrum histogram representation (MCSHR). In [1,15,16], the brightness transfer functions (BTFs) are used to map an observed brightness value in one camera to the corresponding observation in another camera. Once such a mapping is known, the correspondence problem is reduced to the matching of transformed histograms or appearance models. Other work addresses a more general topic for consistent labeling people across (disjoint) camera views called the person re-identification problem [17–20]. In these work, more complex representations are used, such as spatial graph [18], spatial co-occurrence matrix [17,20], boosted features [19]. Due to the computational issue, they are not widely used in tracking; however, color features are still widely used in these work but not too many efforts are taken by [17–20] to alleviate the effect of lighting variations.

1.1.3. Combining spatial–temporal and visual cues

Recently, spatio-temporal and visual cues are combined to infer the correspondence. Gilbert and Bowden [21] combined color information and the posterior probability distribution of spatio-temporal links between cameras for modeling the correspondence function. Javed et al. [22] modeled the inter-camera relationships in terms of multivariate probability density of spatio-temporal variables (entry and exit locations, velocities, and transition times) using kernel density estimation, and handle the appearance change of an object using probabilistic principal component analysis in a low dimensional BTFs subspace for appearance matching. Chen et al. [23] extended these approaches by learning spatio-temporal relationships and BTFs adaptively. In [23], the BTFs are computed for each pair of cameras and used to track individuals across multiple non-overlapping cameras. However, the BTFs are not unique and vary from frame to frame depending on a large number of parameters, including illumination, scene geometry, exposure time, focal length, and aperture size of each camera. To this end, this method needs a lot of data and time to learn the BTFs.

1.2. Our contributions

Though consistent labeling performance for non-overlapping views has been enhanced by incorporation of spatio-temporal and visual cues, existing methods require accurate segmentation of each object or estimation of the trajectories of objects across camera views, and occlusion (by people themselves) is still an issue that has not been too much addressed. Also, robust appearance matching against lighting variations is still an unsolved problem, as lighting conditions would dramatically change between disjoint camera views. To address these problems, we first formulate a hypothesis space for each new object detected in the entry FOV of a camera; and such a hypothesis space of a new detected object consists of the hypotheses that contain a single object or a group objects that leave the exit FOV of another camera in advance. And one of the hypotheses would possibly correspond to that new detected object. A Bayesian framework in the hypothesis space is then established to explore spatio-temporal cue assisted by visual cue to perform the consistent labeling for tracking people across non-overlapping camera views, which can both enhance the performance and cope with appearance and occlusion variations.

Moreover, by using the proposed Bayesian model, if a new object detected in the entry FOV of a camera corresponds to a group hypothesis in which the objects leave the exit FOV of another camera before, the new object will be represented by the ensemble of the labels of the objects in the group hypothesis. In this case, an online update algorithm using optimal graph matching (OGM) is developed to perform the consistent labeling between objects from disjoint camera views.

For alleviating the lighting problem for appearance matching, we extend the major color spectrum histogram representation (MCSHR) [13] and propose the competitive major color spectrum histogram representation (CMCSHR). CMCSHR differs from MCSHR in that the major colors are learned by competitive clustering algorithm in order to not only learn a more robust major colors but also automatically determine the number of them. Compared to the popular BTFs using for consistent labeling across camera views, CMCSHR is much more robust to the change of lighting condition.

Though the Bayesian framework in our work and Calderara's approach [5] are similar, ours differs from Calderara's approach in that the objectives are totally different. Our framework is for consistent labeling across non-overlapping camera view while Calderara's is for overlapping views. Consequently, our Bayesian framework unifies spatial–temporal cue and visual appearance cue, while Calderara's utilizes geometric information. This results in significantly different modeling. Moreover, our proposed model has a first attempt to address the occlusion problem/group problem, i.e. a group of people walking together into the entry FOV of one camera can correctly correspond to the individuals in another camera view.

In summary, this paper makes three main contributions. (1) The Bayesian model is used to perform the consistent labeling across non-overlapping camera views without accurate segmentation for each individual object. (2) A new effective color feature representation namely CMCSHR for appearance matching is presented. (3) An online algorithm using OGM method is proposed to deal with the consistent labeling against the occlusion problem.

The rest of the paper is organized as follows. Section 2 presents our Bayesian model for consistent labeling for tracking pedestrians across non-overlapping camera views. Online update of correspondence using the OGM method is presented in Section 3. Experimental results are described in Section 4. Conclusions and discussions are presented in Section 5.

2. A spatial–temporal Bayesian consistency labeling model

Suppose we have a surveillance system consists of K cameras C^1, $C^2,\ldots,C^K$ with non-overlapping views, and a topological network of the cameras is built to connect cameras based on their temporal extents. Each camera is a node on the network. In our system, we assume that the connected path between each pair of cameras is unique and deterministic. Let t^i_{ex} and t^j_{en} be the exit and entry time of an object moving at normal speed from exiting camera C^i to entering camera C^j, respectively. Let T be a positive temporal threshold. It is roughly the maximum transition time of objects crossing the gap between adjacent cameras. If the time t^i_{ex} and

t^j_{en} satisfy below:

$$t^j_{en} < t^i_{ex} + T, \tag{1}$$

then camera C^i and camera C^j will be connected on the topological network. In most situations, this constraint is reasonable. An example that an object moves at normal speed along the arrow direction can be found in Fig. 1. As shown in (a), the views of cameras are disjoint. Their temporal relationship is as shown in (b). The entry time t^2_{en} of camera C^2 and the exit time t^1_{ex} of camera C^1 satisfy Eq. (1), so camera C^1 and camera C^2 are connected on the network. The gap of t^3_{en} and t^2_{en} is smaller than T, whereas the gap of t^3_{en} and t^1_{en} is larger than T, so camera C^2 and camera C^3 are also connected, but camera C^1 and camera C^3 are not connected on the network.

For the sake of clarity, suppose we have only two non-overlapping cameras C^1 and C^2, and if consistent labeling problem is solved between any pair of cameras, extension to a number of K cameras will be straightforward with a little modification. The proposed approach is based on two non-overlapping camera views as depicted in Fig. 2 and assume that a single based camera tracking algorithm is already available. Let O_{new} be a new object that is detected in the entry field of camera C^2. The consistent labeling is to link the new object O_{new} entering camera C^2 to the subset of N potentially matched objects exiting camera C^1, where it is assumed that the N objects can be detected separately in camera C^1. These N objects must satisfy the following temporal constraints:

$$t^2_{en} - T_{max} \le t^1_{ex} \le t^2_{en} - T_{min}, \tag{2}$$

where t^1_{en} is the exit time of a potentially matched object which exits the FOV of camera C^1, t^2_{en} is the entry time of the new object O_{new} which enters the FOV of camera C^2, T_{max} and T_{min} are the maximum time interval and minimum time interval from exiting camera C^1 to entering camera C^2 for an object, respectively. These N objects are combined to form the hypothesis space $\Gamma = \{\varphi_k | k = 1,2,\ldots,2^N-1\}$, in which each hypothesis includes a single object or a group objects [5]. That is, the new object O_{new} possibly corresponds to a single object or a group objects.

Once the hypothesis space Γ is formulated, similar to [5], a maximum-a-posteriori (MAP) estimator is adopted to find the most probable hypothesis φ_i as follows:

$$i = \arg\max_k p(\varphi_k | O_{new}) = \arg\max_k p(O_{new} | \varphi_k) p(\varphi_k). \tag{3}$$

To compute the MAP, we must estimate the prior of each hypothesis φ_k (i.e. $p(\varphi_k)$) and the likelihood of the new object O_{new} given the hypothesis φ_k (i.e. $p(O_{new}|\varphi_k)$). In our study, the prior computation is modeled with spatio-temporal cue, and a CMCSHR method is proposed to compute the likelihood. Furthermore, we present an online algorithm for update of correspondence using OGM method in order to address the occlusion problem if the detected new object O_{new} corresponds to a group hypothesis (more than one object).

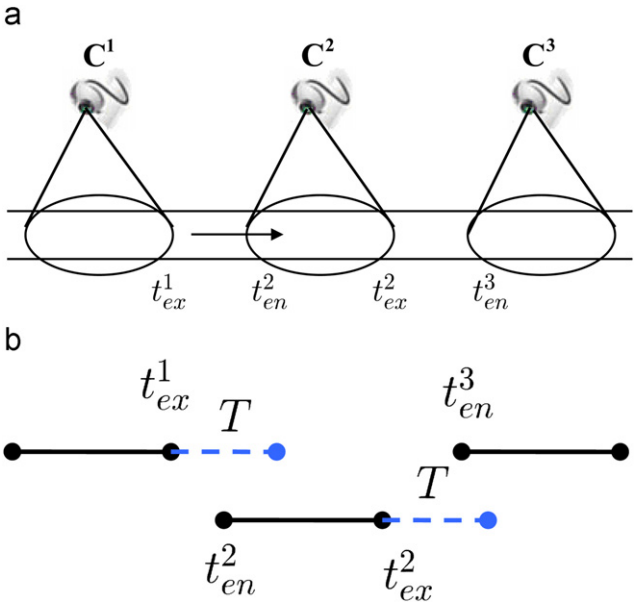

Fig. 1. An example of camera setup and their temporal relationship in multiple non-overlapping camera views: (a) Camera setup. (b) Temporal relationship between camera views. See text for details.

2.1. Prior computation using spatio-temporal cue

The prior of a hypothesis indicates how likely it will happen. Specifically, let $O=\{O_1,O_2,\ldots,O_N\}$ be the set of N objects that exit camera C^1. Then the hypothesis space of these N objects is formulated by $\Gamma = \{\varphi_k | k = 1,2,\ldots,2^N-1\}$ that contains all the (2^N-1) possible subsets, including single object and groups. Since the prior of the new object O_{new} is evaluated, no information about O_{new} must be used. Usually, a uniform distribution in the hypothesis space Γ is used for the prior computation of a given hypothesis φ_k, that is, $p(\varphi_k) = p_0 = 1/|\Gamma|$. However, it is known that only the uniform distribution of the prior cannot provide any information for the selection of MAP in Bayesian decision theory [25]. Intuitively, a hypothesis will gain higher prior value if the objects composing the hypothesis enter the entry FOV of camera C^2 almost

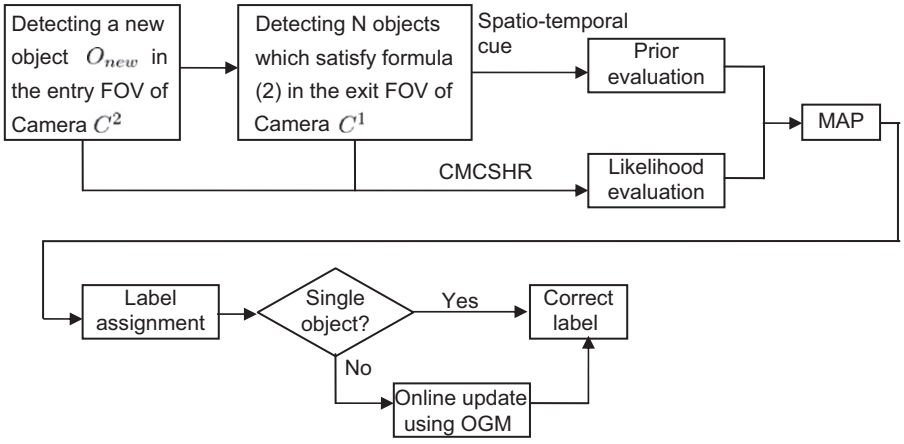

Fig. 2. The Bayesian model approach for consistent labeling across two non-overlapping camera views.

simultaneously (that is these objects can be considered as a whole group), while the other objects in a different hypothesis enter the entry FOV of C^2 at very different time. On the other hand, a hypothesis will gain lower prior if the objects composing the hypothesis enter the entry FOV of camera C^2 at different time but some of them access the entry FOV of camera C^2 with other objects in a different hypothesis.

Based on the above consideration, similar to [24], we model the prior by a combination of a uniform distribution and a probability variant that biases the uniform distribution for each hypothesis according to the spatio-temporal cue. The spatio-temporal cue is modeled based on the constraint in Eq. (2) between the exit time of these N objects and the entry time of O_{new} which is observed in the entry FOV of camera C^2. And also, let t_j^1 be the exit time of camera C^1 and t_j^2 be the possible entry time of camera C^2 when an object O_j moves from camera C^1 to camera C^2. The t_j^2 is evaluated as follows:

$$t_j^2 = t_j^1 + \frac{d}{v_j}, \tag{4}$$

where d is the distance between the exit FOV of camera C^1 and the entry FOV of camera C^2, v_j is the velocity of the object O_j. The time t_j^2 is used for prior computation. The prior of a given hypothesis φ_k is evaluated by assigning a value proportional to a score σ_k [5]. The score σ_k of the hypothesis φ_k is determined by the time difference between objects within φ_k (within-hypothesis time difference) and the time difference between any object in φ_k and any object in other hypotheses (between-hypothesis time difference). Accordingly, the score σ_k of the hypothesis φ_k is computed as the difference between the within-hypothesis time difference and the between-hypothesis time difference. Specifically, the within-hypothesis time difference is computed as follows:

$$W_{td} = \max_{\{O_a,O_b\} \in \varphi_k} |t_a^2 - t_b^2|. \tag{5}$$

The between-hypothesis time difference is computed as follows:

$$B_{td} = \min_{O_a \in \varphi_k, O_b \in \Gamma - \{\varphi_k\}, a \neq b} |t_a^2 - t_b^2|. \tag{6}$$

Then, similar to [24], the relative score associated to each hypothesis is computed by $\sigma_k = B_{td} - W_{td}$. The score σ_k of a given hypothesis φ_k is high, which indicates that the objects composing the hypothesis enter the entry FOV of camera C^2 almost simultaneously, while the other objects in a different hypothesis enter the entry FOV of C^2 at very different time. As discussed in the above description, the hypothesis φ_k with higher score σ_k will gain higher prior. Therefore, we set the prior probability of each hypothesis in the hypothesis space Γ as the combination of the uniform distribution $p_0 = 1/|\Gamma|$ and the probability variant $\Delta\sigma(\varphi_k)$ that is computed by assigning a value proportional to the score σ_k [24]. That is, the prior can be modeled by

$$p(\varphi_k) = p_0 + \Delta\sigma(\varphi_k), \tag{7}$$

with

$$\Delta\sigma(\varphi_k) = \frac{1/(|\Gamma|+1)}{\max_{i=1,\ldots,|\Gamma|}(\sigma_i) - \min_{i=1,\ldots,|\Gamma|}(\sigma_i)} \left(\sigma_k - \frac{\sum_{i=1}^{|\Gamma|} \sigma_i}{|\Gamma|} \right). \tag{8}$$

Note that $p(\varphi_k)$ conforms to the definition of the probability, because the probability variant $\Delta\sigma(\varphi_k)$ satisfies $\Delta\sigma(\varphi_k) < p_0$ and $\sum_{k=1}^{|\Gamma|} \Delta\sigma(\varphi_k) = 0$ (described in Eq. (8)).

As an example shown in Fig. 3, a new object O_{new} is detected in the entry FOV of camera C^2. The four objects (O_{22}, O_{23}, O_{24}, O_{25}) under the temporal constraint of Eq. (2) are retrieved in the exit FOV of camera C^1. The hypothesis space Γ is then formulated as follows:

$$\begin{aligned}\Gamma = \{&\{O_{22}\},\{O_{23}\},\{O_{24}\},\{O_{25}\},\{O_{22},O_{23}\},\{O_{22},O_{24}\},\{O_{22},O_{25}\},\\ &\{O_{23},O_{24}\},\{O_{23},O_{25}\},\{O_{24},O_{25}\},\{O_{22},O_{23},O_{24}\},\{O_{22},O_{23},O_{25}\},\\ &\{O_{22},O_{24},O_{25}\},\{O_{23},O_{24},O_{25}\},\{O_{22},O_{23},O_{24},O_{25}\}\}.\end{aligned}$$

Using the previously described prior model, the values reported in Table 1 are obtained.

Table 1
An example of computing σ_k and the prior of each hypothesis.

Hypothesis	σ_k value	Prior value
$\varphi_1 = \{O_{22}\}$	2.2	0.0882
$\varphi_2 = \{O_{23}\}$	1.0	0.0821
$\varphi_3 = \{O_{24}\}$	1.0	0.0821
$\varphi_4 = \{O_{25}\}$	4.5	0.1000
$\varphi_5 = \{O_{22},O_{23}\}$	−1.2	0.0708
$\varphi_6 = \{O_{22},O_{24}\}$	−2.2	0.0657
$\varphi_7 = \{O_{22},O_{25}\}$	−5.5	0.0488
$\varphi_8 = \{O_{23},O_{24}\}$	1.2	0.0831
$\varphi_9 = \{O_{23},O_{25}\}$	−4.5	0.0539
$\varphi_{10} = \{O_{24},O_{25}\}$	−3.5	0.0590
$\varphi_{11} = \{O_{22},O_{23},O_{24}\}$	1.3	0.0836
$\varphi_{12} = \{O_{22},O_{23},O_{25}\}$	−6.7	0.0426
$\varphi_{13} = \{O_{22},O_{24},O_{25}\}$	−6.7	0.0426
$\varphi_{14} = \{O_{23},O_{24},O_{25}\}$	−3.3	0.0600
$\varphi_{15} = \{O_{22},O_{23},O_{24},O_{25}\}$	−7.7	0.0375

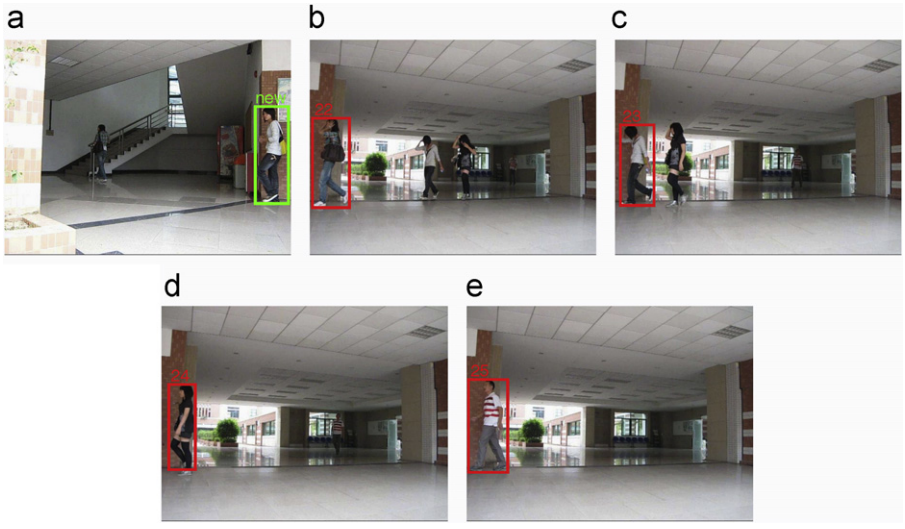

Fig. 3. An example of the detection events occurring in camera C^1 and C^2. (a) A new object was detected in the entry FOV of camera C^2. (b)–(e) Some objects have been detected in the exit FOV of camera C^1. And the time of those objects exiting camera C^1 and the time of the new object entered camera C^2 satisfied Eq. (2).

2.2. Likelihood computation based on CMCSHR

The likelihood $p(O_{new}|\varphi_k)$ indicates the probability of the existence of O_{new} in camera C^2 if the hypothesis φ_k is held. Since we are dealing with the tracking problem across non-overlapping camera views, geometric information which is always adopted for the overlapping case is not suitable here [24,5]. We then wish to use the appearance information and particularly focus on color information which is the most useful and effective feature for the tracking problem. However, color is always largely affected by lighting. Recently, Madden et al. [13] introduced the Major Color Spectrum Histogram Representation (MCSHR) for tracking people across disjoint camera views. In their method, an illumination-tolerant appearance representation is proposed, which is capable of coping with the typical illumination changes occurring in surveillance scenarios, but its weakness is that the performance is greatly affected by the amount of major colors should be exploited and the initial set of major color clusters which is critically important. And it is also computationally expensive to obtain the major colors by MCSHR. In view of this, we modify MCSHR and propose the Competitive Major Color Spectrum Histogram Representation (CMCSHR) in which competitive clustering technique is utilized in order to efficiently compute more stable major color and simultaneously determine the amount of many major colors should be set. In the following, we first introduce MCSHR in brief, then describe the proposed CMCSHR, and finally present the likelihood computation.

2.2.1. MCSHR

In the Major Color Spectrum Histogram Representation (MCSHR) introduced by Madden et al. [13], a distance between any two color pixels is firstly defined in the RGB space. The major colors of an object are obtained based on an online k-means clustering algorithm.

Suppose there exists M major colors in object A which can be represented as:

$$MCSHR(A)=\{C_{A_1},C_{A_2},\ldots,C_{A_M}\}, \tag{9}$$

with their frequencies calculated by normalizing the major colors histogram as follows:

$$p(A)=\{p(A_1),p(A_2),\ldots,p(A_M)\}. \tag{10}$$

Similarly, object B can be represented over N major colors by the $MCSHR(B)=\{C_{B_1},C_{B_2},\ldots,C_{B_N}\}$ and $p(B)=\{p(B_1),p(B_2),\ldots,p(B_N)\}$. In order to define the similarity between two objects, for each C_{A_i}, the most similar color in object B is denoted as $C_{B_j|A_i}$ with the similar constraint $d(C_{B_j|A_i},C_{A_i})<\lambda$. To obtain the $C_{B_j|A_i}$ in object B, a subset of $MCSHR(B)$ of object B in which the colors are considered to be close enough to C_{A_i} is firstly defined as

$$MCSHR'(B)=\{C_{B'_1},C_{B'_2},\ldots,C_{B'_K}\}, \tag{11}$$

with $C_{B'_k}\in MCSHR(B)$ and $d(C_{B'_k},C_{A_i})<\lambda, k=1,2,\ldots,K$, where λ is a given threshold. Accordingly, the frequency set of the major colors in $MCSHR'(B)$ is $\{p(B'_1),p(B'_2),\ldots,p(B'_K)\}$, which is a subset of $\{p(B_1),p(B_2),\ldots,p(B_N)\}$. Then, $C_{B_j|A_i}$ is defined as

$$C_{B_j|A_i}: j=\underset{k=1,\ldots,K}{\operatorname{argmin}}\{d(C_{B'_k},C_{A_i})\}. \tag{12}$$

The frequency of major color $C_{B_j|A_i}$ in object B is computed as follows:

$$p(B_j|A_i)=p(B'_j). \tag{13}$$

The similarity between object A and object B in the direction from A to B is then given by

$$\begin{aligned} Sim(A,B)&=\sum_{i=1}^{M} Sim(C_{A_i},C_{B_j|A_i})\\ &=\sum_{i=1}^{M} min\{p(A_i),p(B_j|A_i)\}. \end{aligned} \tag{14}$$

Because histogram intersection [26] is not scale invariant, the major color histograms of objects should be scaled/normalized to be the same size.

2.2.2. Proposed CMCSHR

In the RGB space, there are about 16.8 million colors. It is very difficult to compare two objects based on so many possible colors. Although Madden et al. [13] introduced a color distance, the color distance between red and green is the same as the one between red and blue. And the brightness of the same object is the major difference between two non-overlapping views. To overcome these drawbacks, we use the hue and saturation as the feature vectors. Also, the performance of MCSHR is greatly affected by the initial set of major color clusters. It would be an important issue if we do not have enough prior knowledge on it. In order to automatically determine the number of major colors, we develop a Competitive Learning based Major Color Spectrum Histogram Representation (CMCSHR). The rest part of this section then details these two developments.

HS-distance. Hue which is invariant to brightness for a given RGB color space is formulated as follows [18]:

$$H=\arccos\frac{\log(R)-\log(G)}{\log(R)+\log(G)-2\log(B)} \tag{15}$$

and saturation is as follows [27]:

$$S=1-\frac{3}{(R+G+B)}[\min(R,G,B)]. \tag{16}$$

Then we normalize H and S and compute the color distance [13] as follows:

$$d(C_1,C_2)=\frac{\sqrt{(H_1-H_2)^2+(S_1-S_2)^2}}{\sqrt{H_1^2+S_1^2}+\sqrt{H_2^2+S_2^2}}, \tag{17}$$

where C_1 and C_2 represent the hue and saturation vectors for the two pixels.

Competitive major colors. The k-means clustering algorithm [28] would lead to a poor clustering performance if the exact number of clusters is not properly set. In our experiments, it is also hard (if not impossible) to know the exact color number of an object. And also, different objects would have different color numbers. There are many clustering algorithms [29–32] that can automatically select the correct cluster numbers, and compared to other algorithms, the Rival Penalization Controlled Competitive Learning (RPCCL) clustering algorithm [29] can converge very fast. For a surveillance system, real-time tracking is very important. In view of this, we use the RPCCL algorithm for color clustering. In the following, we will detail the iterative procedure (steps 1 and 2) for learning competitive major colors using RPCCL and a short introduction of RPCCL is also given in Appendix A.

Step 1: Randomly take k pixels from the object as the seed points, denoted as $\{m_j\}_{j=1}^{k}$, and the object's pixels are scanned in row-major order. For each pixel $p_i=(p_i^H,p_i^S)^T$ from the object, where p_i^H and p_i^S are the hue and saturation value of the pixel p_i, respectively, we define its role indicator as follows:

$$I(j|p_i)=\begin{cases}1 & \text{if } j=c(p_i),\\ -1 & \text{if } j=r(p_i),\\ 0 & \text{otherwise,}\end{cases} \tag{18}$$

with

$$\begin{aligned} c(p_i)&=\underset{j}{\operatorname{argmin}}\gamma_j d(p_i,m_j)\\ &=\underset{j}{\operatorname{argmin}}\gamma_j\frac{\sqrt{(p_i^H-m_j^H)^2+(p_i^S-m_j^S)^2}}{\sqrt{(p_i^H)^2+(p_i^S)^2}+\sqrt{(m_j^H)^2+(m_j^S)^2}}, \end{aligned} \tag{19}$$

$$r(p_i) = \underset{j \neq c(p_i)}{\operatorname{argmin}} \gamma_j d(p_i, m_j) = \underset{j \neq c(p_i)}{\operatorname{argmin}} \gamma_j \frac{\sqrt{(p_i^H - m_j^H)^2 + (p_i^S - m_j^S)^2}}{\sqrt{(p_i^H)^2 + (p_i^S)^2} + \sqrt{(m_j^H)^2 + (m_j^S)^2}}, \quad (20)$$

where $\gamma_j = n_j / \sum_{r=1}^{k} n_r$ is the relative winning frequency of the seed point m_j in the past, and n_j is the cumulative number of the occurrences of $I(j|p_i) = 1$ in the past.

Step 2: Update the winner m_c(*i.e.*, $I(c|p_i) = 1$) and its rival m_r(*i.e.*, $I(r|p_i) = -1$) only by

$$m_u^{new} = m_u^{old} + \Delta m_u,\ u = c, r, \quad (21)$$

with

$$\Delta m_c = \alpha_c (p_i - m_c), \quad (22)$$

$$\Delta m_r = -\alpha_c p_r(p_i)(p_i - m_r), \quad (23)$$

$$p_r(p_i) = \frac{\min(|m_c - m_r|, |m_c - p_i|)}{|m_c - m_r|}, \quad (24)$$

where α_c is the learning rate. For this pixel, the closest cluster center m_c is computed and the pixel assigned to it. Since cluster centers $\{m_j\}_{j=1}^{k}$ are moved, iterations are necessary until all pixel assignments and cluster centers stabilize. We call the set of these clustered colors the competitive major colors.

2.2.3. Likelihood computation

After learning competitive major colors using CMCSHR, we first adopt the similarity metric as described in [13] to compute the distance between two objects. Accordingly, the fitness measure $\delta_{O_k \to O_{new}}$ from the object O_k in camera C^1 to O_{new} in camera C^2 is defined as the similarity of object O_k and object O_{new} in the direction from O_k to O_{new}, that is,

$$\begin{aligned} \delta_{O_k \to O_{new}} &= Sim(O_k, O_{new}) \\ &= \sum_{i=1}^{M} Sim(C_{O_{ki}}, C_{O_{newj}|O_{ki}}) \\ &= \sum_{i=1}^{M} \min\{p(O_{ki}), p(O_{newj}|O_{ki})\}, \end{aligned} \quad (25)$$

with

$$C_{O_{newj}|O_{ki}} : j = \underset{n}{\operatorname{argmin}} \{d(C_{O_{newn}}, C_{O_{ki}}) | d(C_{O_{newn}}, C_{O_{ki}}) < \lambda\}, \quad (26)$$

where $C_{O_{ki}}$ and $p(O_{ki})$ are the ith major color and its frequency in object O_k, $C_{O_{newn}}$ is the nth major color in object O_{new} and $C_{O_{newj}|O_{ki}}$ denotes the jth major color in object O_{new} is the most similar color to $C_{O_{ki}}$ given a threshold λ, $p(O_{newj}|O_{ki})$ is the frequency of $C_{O_{newj}|O_{ki}}$.

Similarly, the reversed fitness measure $\delta_{O_{new} \to O_k}$ is defined as the similarity between object O_{new} and object O_k in the direction from O_{new} to O_k, that is,

$$\begin{aligned} \delta_{O_{new} \to O_k} &= Sim(O_{new}, O_k) \\ &= \sum_{i=1}^{M} Sim(C_{O_{newi}}, C_{O_{kj}|O_{newi}}) \\ &= \sum_{i=1}^{M} \min\{p(O_{newi}), p(O_{kj}|O_{newi})\}. \end{aligned} \quad (27)$$

Then for the likelihood (similarity/confidence measurement) computation, we adopt the forward and backward contributions as described in [24]. The forward contribution is calculated by computing the fitness measure (described in Eq. (25)) from all the objects composing the given hypothesis φ_k in camera C^1 to the new object O_{new} in camera C^2:

$$p_{forward}(O_{new}|\varphi_k) \propto \sum_{O_m \in \varphi_k} \delta_{O_m \to O_{new}}. \quad (28)$$

Similarly, the backward contribution is computed by the reversed fitness measurement (described in Eq. (27)) from the new object O_{new} in camera C^2 to all the objects composing the given hypothesis φ_k in camera C^1:

$$p_{backward}(O_{new}|\varphi_k) \propto \sum_{O_m \in \varphi_k} \delta_{O_{new} \to O_m}. \quad (29)$$

In order to obtain a more accurate matching, we adopt the strategy in [24] to compute the likelihood by $p(O_{new}|\varphi_k) = \max\{p_{forward}, p_{backward}\}$, so that a more accurate similarity is used for matching.

3. An online update of correspondence using optimal graph matching

If the detected new object O_{new} in the entry FOV of camera C^2 corresponds to a single object in another camera C^1, we naturally perform the consistent labeling between the two objects. However, if O_{new} in camera C^2 corresponds to a group hypothesis which contains $m (m > 1)$ objects in camera C^1, the O_{new} is assigned the ensemble of the labels of these m objects associated to the group hypothesis. In this case, we will track O_{new} in camera C^2 until it splits into $n (n > 1)$ objects. In our experiment, the Particle Filter [33] is used to perform tracking in a single camera view. Then, the n split objects can be labeled with the labels of the m objects using the optimal graph matching (OGM) algorithm. Fig. 4 shows the online update process using OGM.

OGM is a classical problem in graph theory. Let $G = \{X, Y, E\}$ be a bipartite graph, where $X = \{x_1, x_2, \ldots, x_n\}$ denotes the set of the objects in camera C^2 after the group object O_{new} splits, $Y = \{y_1, y_2, \ldots, y_m\}$ is the set of the objects associated to the group hypothesis of O_{new} in camera C^1, $V = X \cup Y$ is the vertex set, and $E = \{e_{ij}\}$ is the edge set. The weight w_{ij} of the edge e_{ij} measures the similarity between object x_i and object y_j. The similarity between two objects is calculated by the proposed CMCSHR. A matching M of G is a subset of the edges with the property that no two edges of M share the same nodes and OGM is to find the matching M that has the largest total weight [34]. Given the weighted bipartite graph G, the Kuhn–Munkres algorithm [35] is employed to solve the OGM problem.

In practice, the scenario of group split can be complicated sometimes. In the following, we further discuss two cases, which are definitely challenging scenarios, and provide our initial solution to these hard problems.

Scenario 1. In our system, after the group object O_{new} splits into n objects, we will first check whether m is equal to n. If it is, the consistent labeling is established between the n split objects in camera C^2 and the m objects in camera C^1.

Sometimes, the group object O_{new} which consists of m objects does not completely split into the exact m objects, that is, $n < m$ after the splitting. Let $M = \{e_{1j_1}, e_{2j_2}, \ldots, e_{nj_n}\}$ denote the matched edge set which is obtained using the OGM algorithm on the graph $G = \{X, Y, E\}$, where e_{ij_i} represents the edge between the split object x_i in C^2 and the single object y_{j_i} in C^1 in graph G. $Y_1 = \{y_{j_1}, y_{j_2}, \ldots, y_{j_n}\}$ is the subset of the vertex set Y and represents the n matched objects in C^1. In this case, there are still $m - n$ single objects in C^1 which are not matched, that is, there must exist some split objects in C^2 still containing more than one object. In order to identify the group objects after the splitting of group object O_{new}, we propose to perform the following operations:

- We first remove the matched vertex subset Y_1 in vertex set Y; that is, $Y' = Y - Y_1$, and we have $G' = \{X, Y', E'\}$, where E' is the subset of the edge set E, and also is the edge set of the graph G' after deleting the Y_1.

- Next, in the graph G', we select the top $m-n$ largest weight value edges to determine which split objects are still group objects and relabel them as group.

The above operations are based on the assumption that the similarity between correct matched non-occluded objects is larger than the one between a non-occluded individual and a group object in which the individual is involved and may be partial-occluded, and the similarity value of the latter case is larger than the similarity values of other matching cases. Based on this assumption, the first matched n objects in C^1 should contain all individuals and some objects in group objects after split. By removing these first n matched nodes in the vertex set Y in G, the next highest matched $m-n$ objects are likely to be part of each group object amid the n split patterns, and therefore the group objects after the splitting can be located. Since each of them is still a group object, we go on using our online matching algorithm to handle it.

Fig. 5 illustrates a synthetic example of the consistent labeling process. In this figure, the split objects x_1 and x_2 are, respectively, labeled as 1 and 4 using OGM algorithm after the first splitting. Since $n(=2)$ is less than $m(=4)$, there must exist some split objects (here x_1 and x_2) still containing more than one object. We then remove the matched vertex subset $Y_1=\{y_1,y_4\}$ (the two vertices have been matched) in the graph G and then we have $G'=\{X,Y',E'\}$ as the graph after the operation. Next, we select the top $m-n$ $(4-2=2)$ largest weight value edges (the edges with weight value 0.7) in the graph G', and x_1, x_2 are relabeled as {1,2} and {3,4}, respectively. After the second splitting, the consistent labeling is established between the $n=4$ split objects and the $m=4$ objects.

A real example of the online update of correspondence using OGM between the $m=2$ single objects of the group hypothesis in camera C^1 and the $n=2$ objects obtained after O_{new}'s splitting in camera C^2 is illustrated in Fig. 6. Other real examples in which O_{new} is assigned the ensemble of the labels of three single objects of the hypothesis in camera C^1 are illustrated in Figs. 7 and 8. Furthermore, the consistent labeling process is shown in Fig. 7(e) and (f).

Scenario 2. If O_{new} which corresponds to a group hypothesis in camera C^1 never splits in camera C^2, O_{new} is assigned the ensemble of the labels of the m objects in the group hypothesis in the next tracking process. When this O_{new} exits camera C^2, the m objects in camera C^1 are considered to simultaneously exit camera C^2, and the system keeps the record of the m individual objects in the group object O_{new}. After O_{new} exits camera C^2, if O_{new} splits between C^2 and C^3 and the split objects are detected in the entry FOV of camera C^3, the hypothesis space of each of the split objects is formed by the m objects in O_{new} and any other objects if they satisfy the temporal constraint in Eq. (2) with the object O_{new}. Fig. 9 shows an example in which O_{new} is assigned the ensemble of the labels of three single objects of the hypothesis in camera C^1. In Fig. 9(b), two split objects are obtained after the first splitting and one of them is labeled with {59,61}. In Fig. 9(c), the split object with label {59,61} never splits again when it exits this camera. In this case, the objects with label 59 and 61 in camera C^1 are considered to simultaneously exit camera C^2. And the objects with label 59 and 61 in camera C^1 are used to perform the consistent labeling between camera C^2 and C^3.

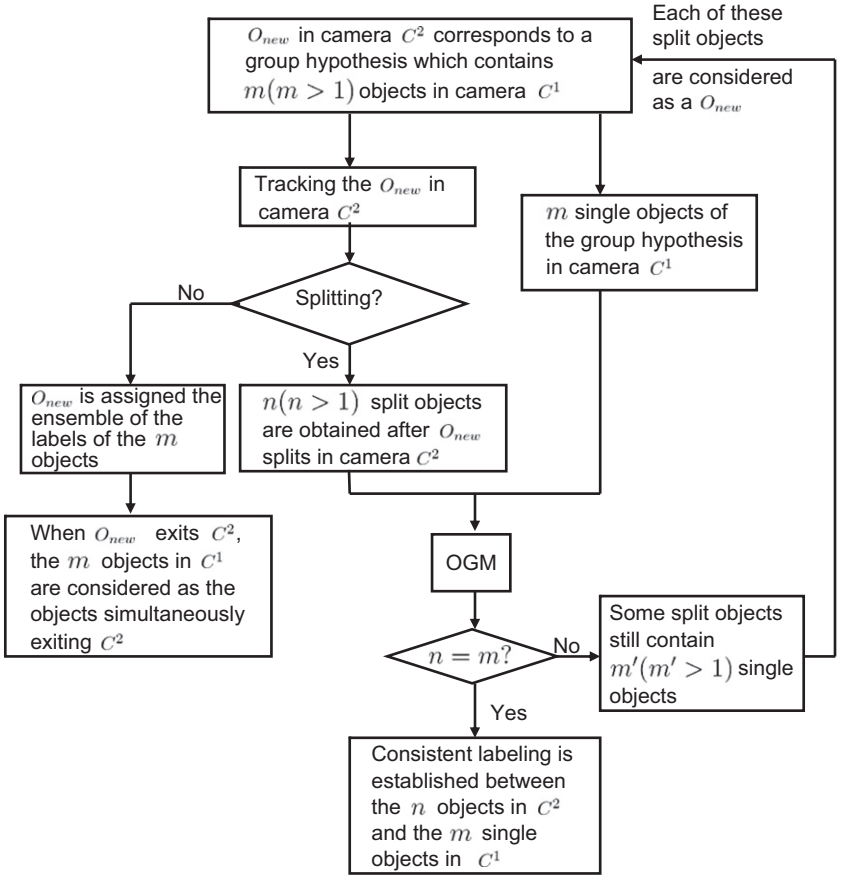

Fig. 4. Illustration of online update of correspondence using OGM.

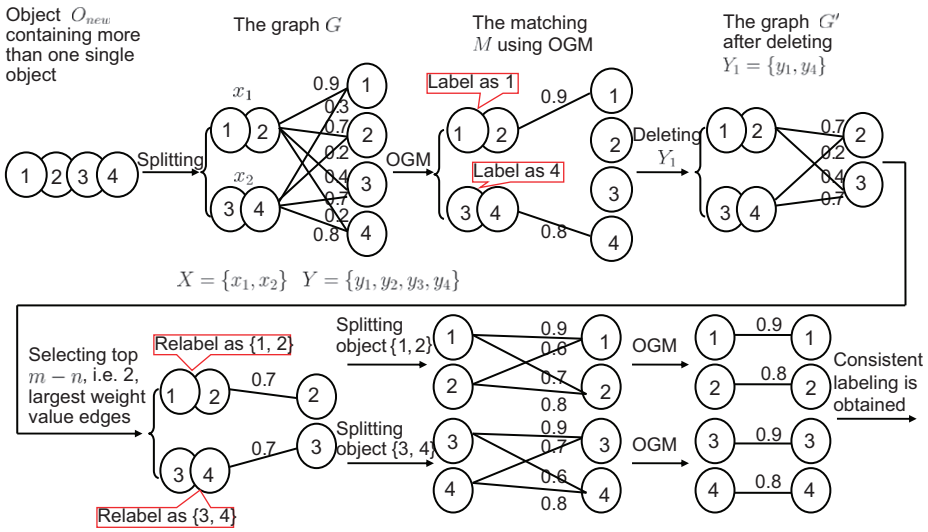

Fig. 5. A synthetic example of the consistent labeling process when the number of the objects after splitting in camera C^2 is less than the number of the objects in the group in camera C^1.

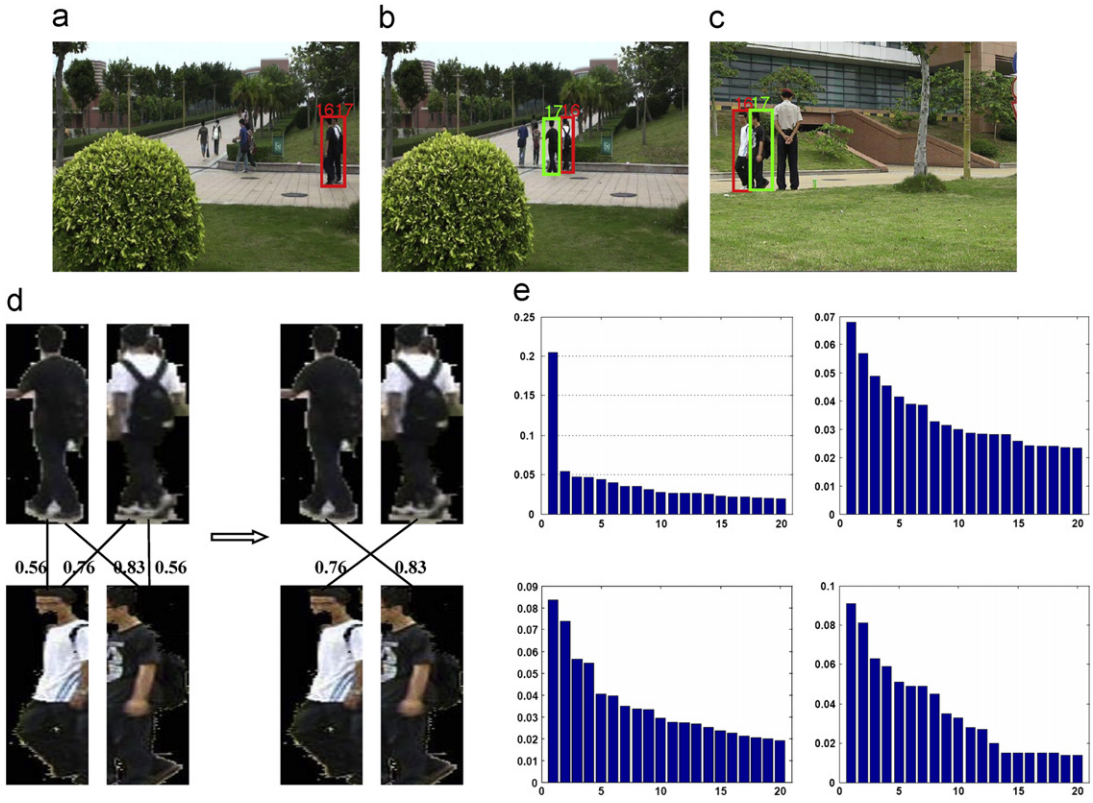

Fig. 6. (a) O_{new} detected in the entry FOV of camera C^2. It was a group hypothesis assigned label {16,17}. (b) O_{new} was split into two objects which were labeled with 16 and 17, respectively. (c) Two objects with label 16 and 17 were detected in the exit FOV of camera C^1. (d) The correspondence was achieved by using CMCSHR and OGM. (e) The CMCSHR of the objects in (d).

4. Experimental results

4.1. Experiment settings

In this section, we report the experimental results of the proposed approach in three different disjoint multi-camera view scenarios. In each experiment, the single camera object detection is based on the background subtraction proposed in [36] and object tracking from each camera is based on the Particle Filter as described in [33]. In the training phase, the known correspondence information is used to compute the spatio-temporal features (inter-camera distance and inter-camera time interval distribution). In the testing phase, the correspondences are computed using the proposed approach.

The three disjoint multi-camera view scenarios differ from each other in terms of distance between two cameras, scene

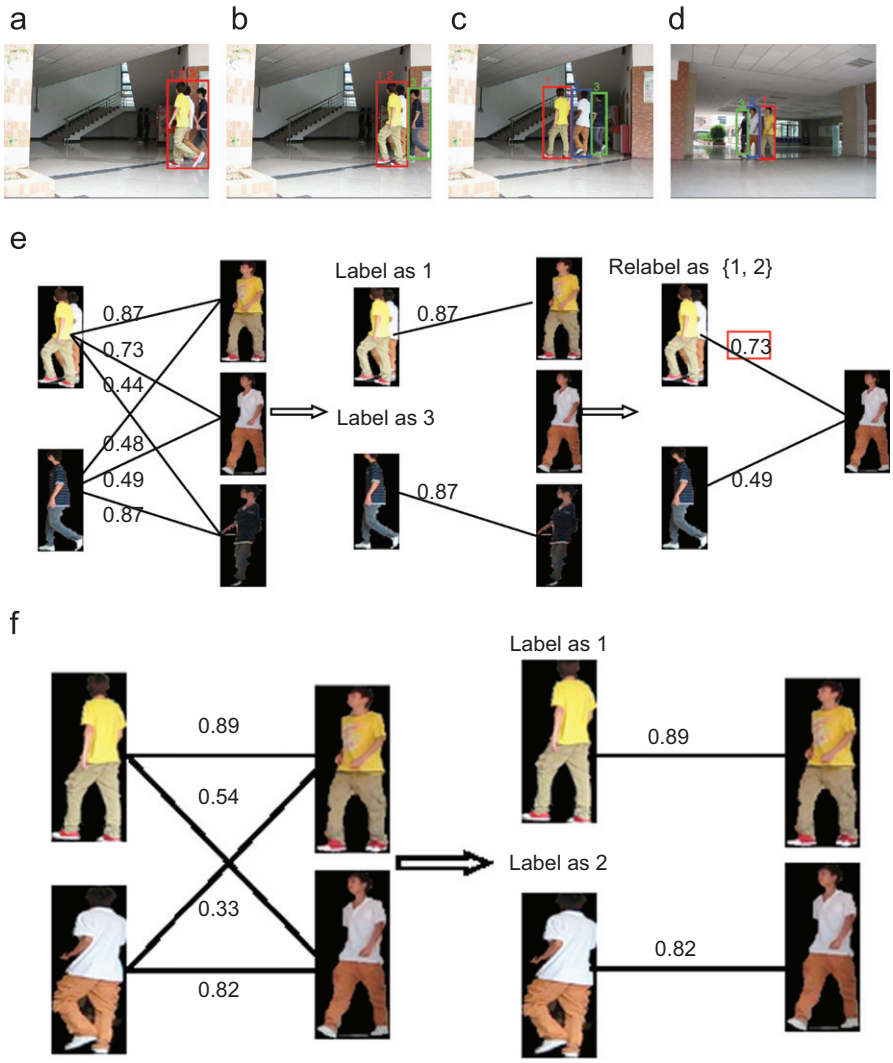

Fig. 7. (a) A new object was detected in the entry FOV of camera C^2. It was a group object consisted of three people. (b) O_{new} split into two objects which were labeled with 3 and {1,2}, respectively. (c) The split object with label {1,2} again split into two objects which were labeled with 1 and 2, respectively. (d) Three objects exiting camera C^1 were obtained. (e) The consistent labeling process is shown after the first splitting. (f) The consistent labeling was performed after the object with label {1,2} split.

illumination conditions, and the environment setting such as indoor and outdoor settings. In scenario 1, the distance between two cameras is large, while in scenario 3, illumination conditions are very different in the two cameras. They are introduced in the following and Table 2 gives a brief summation.

Scenario 1. Experiment on scenario 1 was conducted with two cameras, namely camera C^1 and camera C^2, in an outdoor setting. The camera topology is shown in Fig. 10(a). Training was performed on a 30 min sequence. The cameras were mounted approximately 40 m apart. The inter-camera time interval distribution for a person by walking from exiting camera C^1 to entering camera C^2 is shown in Fig. 10(b). It shows that the time interval from exiting camera C^1 to entering camera C^2 is almost between 25 and 45 s, and most people take about 35 s to walk through the distance. Fig. 11 shows some tracking instances for the testing sequence in scenario 1. In this phase, a total of 36 transitions across the cameras were recorded.

Scenario 2. It consists of two cameras, namely camera C^1 and camera C^2, as shown in Fig. 12. Training was done on a 24 min sequence. In this scenario (scenario 2), the cameras were mounted approximately 23 m apart. The time interval from exiting camera C^1 to entering camera C^2 is between 16 and 21 s, and most people take about 18 s to walk through the distance. Fig. 13 shows some tracking instances for a testing sequence in which 46 transitions were detected across the cameras. Another testing sequence was captured at the different time. Fig. 8 shows an instance of this sequence. In this sequence, 16 transitions were detected across the cameras. Hence, a total of 62 transitions for the two testing sequences were detected across the cameras in this scenario.

Scenario 3. In scenario 3, two cameras C^1 and C^2 were used for an indoor/outdoor setup. Camera C^1 was placed outdoor while camera C^2 was placed indoor. The placement of the cameras along with their fields of view is shown in Fig. 14. The scene viewed by camera C^1 is a hall in which it is a covered area under shade, whereas camera C^2 monitored the area that is very dim (as shown in Fig. 15). It can be seen from Fig. 15 that there is a significant difference between the global illumination of the two scenes. Training was done on a 25 min sequence. In this scenario, the cameras were mounted approximately 23 m apart. The time interval from exiting camera C^1 to entering camera C^2 is between 16 and 21 s for people to walk through the distance. Fig. 15 shows some tracking instances for the testing sequence in scenario 3. From Fig. 15, it can be seen that two people with the same color clothes from exiting camera C^1

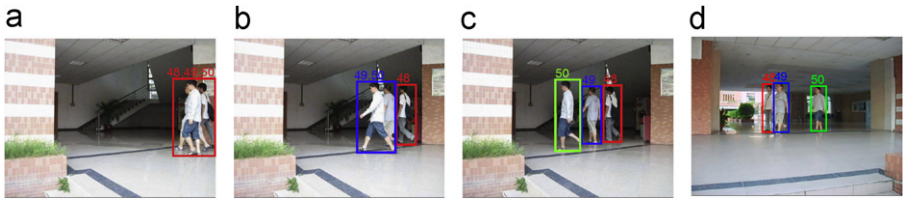

Fig. 8. (a) O_{new} detected in the entry FOV of camera C^2. It was a group hypothesis assigned the ensemble of three labels. (b) O_{new} split into two objects which were labeled with 48 and {49,50}, respectively. (c) The split object with label {49,50} again split into two objects which were labeled with 49 and 50, respectively. (d) Three objects exiting camera C^1 were obtained.

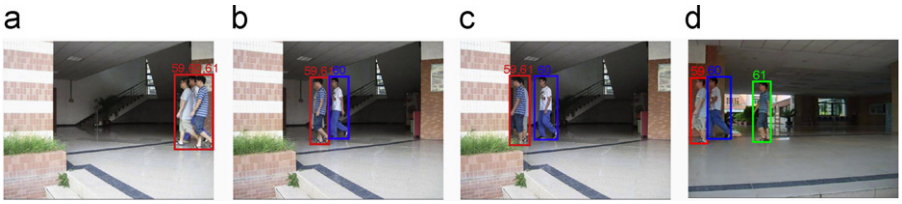

Fig. 9. (a) O_{new} detected in the entry FOV of camera C^2. It was a group hypothesis assigned the ensemble of three labels {59,60,61}. (b) O_{new} split into two objects which were labeled with 60 and {59,61}, respectively. (c) The split object with label {59,61} does not split when it exited this camera. (d) Three objects exiting camera C^1 were obtained.

Table 2
Minimum time interval (T_{min}), maximum time interval (T_{max}) estimated from training data, and total transitions used to test different methods.

	Distance between two cameras (m)	T_{min} (s)	T_{max} (s)	Total transitions (#)
Scenario 1	40	25	45	36
Scenario 2	23	16	21	62
Scenario 3	23	16	21	39

to entering camera C^2 are assigned correct labels. In this testing phase, a total of 39 transitions across the cameras were recorded.

4.2. Results

4.2.1. Evaluation of the proposed model

Among the different testing sequences analyzed, all transitions have been manually annotated to obtain the ground truth. By observing the three actual scenarios in the testing phase, the results from our proposed Bayesian model are reported in Table 3. A visual summary of these results is shown in Fig. 16. Moreover, we analyze the effects of different ingredients in our model, namely the time information, the proposed CMCSHR and the Bayesian model. As shown in Fig. 16, we compare (i) only using time model, (ii) only using the proposed CMCSHR, and (iii) the proposed Bayesian model (Criterion (3)). It shows that the combination of spatio-temporal and appearance information using our Bayesian model is better than only using one of them for all scenarios.

In our proposed model, the consistent labeling is to link the new object O_{new} detected in the entry FOV of camera C^2 to the subset of N potentially matched objects exiting the exit FOV of camera C^1. The N objects must satisfy the temporal constraint in Eq. (2). However, if a person passes across two cameras at an unusual speed, such as running or walking in a very slow speed, the consistent labeling operation will fail. Fig. 17 shows an incorrect correspondence example in which a person ran across the two cameras.

We also further evaluate the proposed CMCSHR by comparing it with MCSHR and BTF.

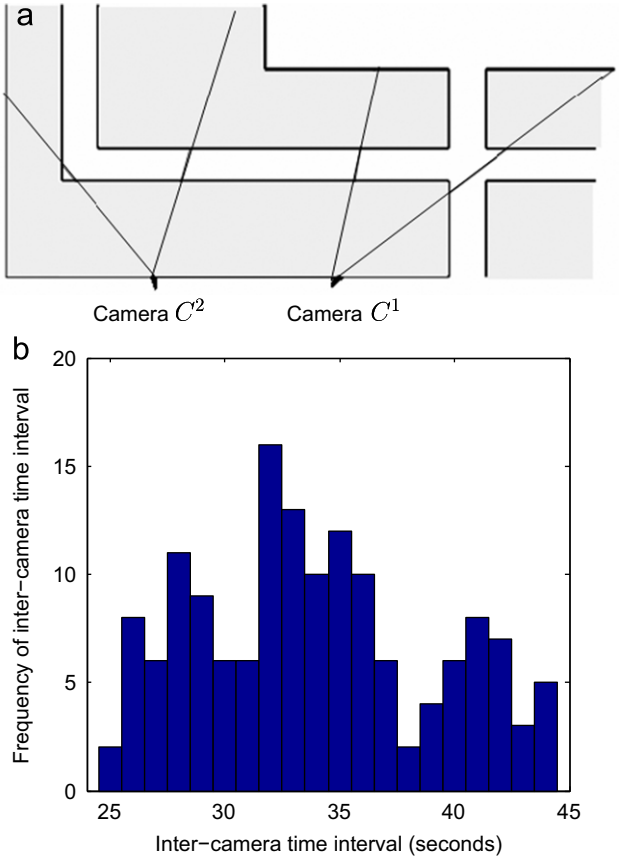

Fig. 10. (a) Camera configuration for the first experiment (scenario 1). (b) The histogram of the inter-camera time interval (learned from training data). Note that most people take almost 35 s to walk from camera C^1 to camera C^2; the minimum time is about 25 s and the maximum time is about 45 s.

CMCSHR vs. MCSHR: To demonstrate the superiority of the proposed CMCSHR over MCSHR, Fig. 18 shows the comparison between only using CMCSHR and only using MCSHR. The results

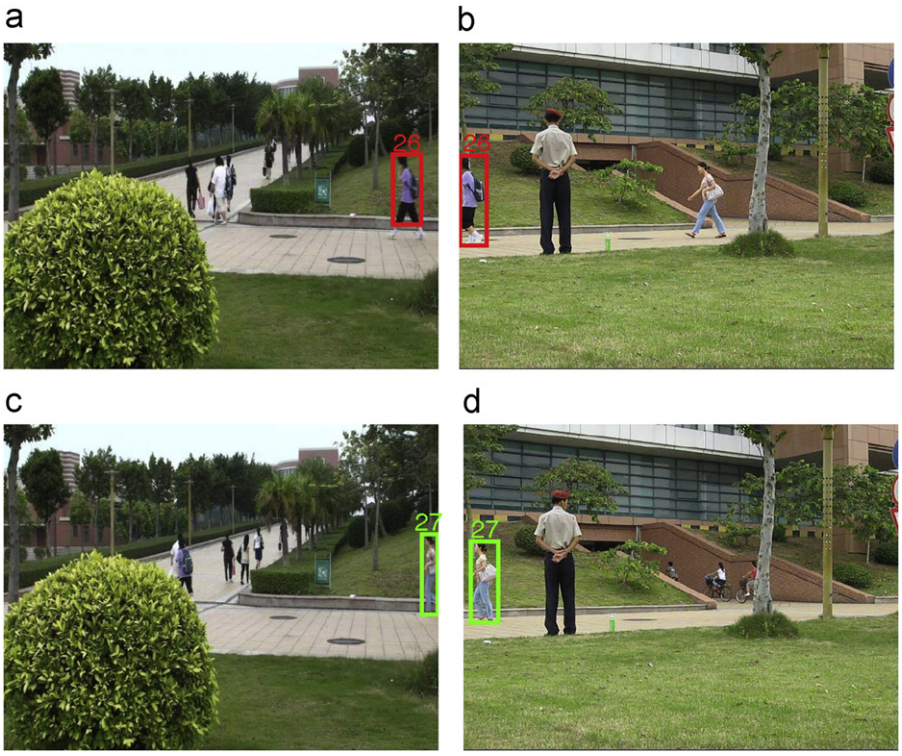

Fig. 11. Frames extracted from the testing phase of scenario 1. A person was assigned a correct label as it moved across two camera views. A person was detected in the entry FOV of camera C^2 in (a) and (c), and people which satisfied Eq. (2) have been detected in the exit FOV of camera C^1 in (b) and (d).

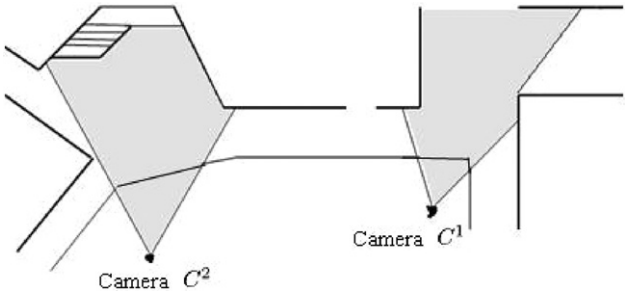

Fig. 12. Camera setup for scenario 2.

show that CMCSHR obtains around 6% more correct matching rate. In order to show the advantages of using hue and saturation as features rather than RGB features and using competitive clustering technique rather than Kmeans, we compare RGB/HS+Kmeans with RGB/HS+RPCCL within the framework of our proposed Bayesian model in all three scenarios. As shown in Fig. 19, CMCSHR performs the best, and this suggests that the effect of object appearance variation is reduced by using hue and saturation as features and more effective major color clusters are estimated by using RPCCL clustering algorithm.

CMCSHR vs. CBTF: In order to show that our proposed CMCSHR method is effective, we compare the CMCSHR method with the Cumulative Brightness Transfer Function (CBTF) approach [15]. In [15], instead of computing a BTF for each pair of training objects, an accumulation of the brightness values of the whole training set is obtained before the BTF computation. Fig. 20 shows that CMCSHR obtains an approximate 7% matching rate improvement as compared to the CBTF in all scenarios.

4.2.2. Proposed Bayesian model vs. Javed's model

To show our proposed Bayesian model is more effective in using spatio-temporal and appearance information for tracking people across non-overlapping camera views, we compare our Bayesian model with a most related work namely Javed's model approach [22]. In [22], the spatio-temporal and appearance cues are integrated to constrain correspondences using BTFs, which is similar to ours. Fig. 21 shows the BTFs taken from our three real scenarios between the exit FOV of camera C^1 and the entry FOV of camera C^2. The comparison results are presented in Fig. 22, and clearly show that our Bayesian model method performs better. Besides, as shown later, our method can deal with the consistent labeling under occlusions which occurs in the entry FOV of camera C^2. However, Javed's approach needs accurate segmentation of the objects into individuals before it performs the consistent labeling between camera views. For example, as shown later, when a new object consisting of two people was detected in the entry FOV of camera C^2, the new object can correctly correspond to a hypothesis which contains the two individual people detected in the exit FOV of camera C^1 using our Bayesian model method; while, in this case, Javed's approach cannot make any correspondence between the new object and the two people.

4.2.3. Consistent labeling against occlusions

Theoretically, our proposed method can deal with the occlusion which occurs a few people entering the entry FOV of camera C^2 almost simultaneously. In Eq. (3), a MAP estimator is adopted to find the most probable hypothesis φ_i. Two or more people who are occluded each other in the entry FOV of camera C^2 are treated as a detected new object. According to the prior and likelihood computation, a hypothesis formed by all these people who compose the new object should have higher prior and likelihood. So this hypothesis is the most probable hypothesis. To demonstrate that our proposed method can deal with the occlusions, we conducted

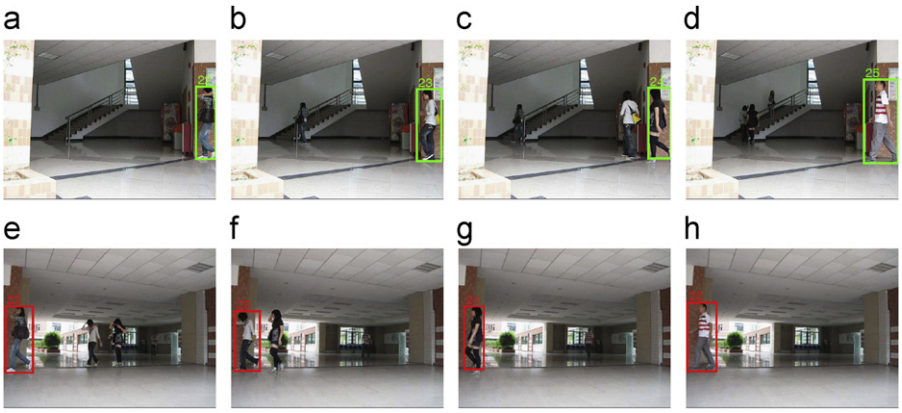

Fig. 13. Frames extracted from the testing phase of scenario 2. A person was assigned a correct label as it moved across two camera views. The first row shows that a person was detected in the entry FOV of camera C^2, the second row shows that people which satisfied Eq. (2) have been detected in the exit FOV of camera C^1.

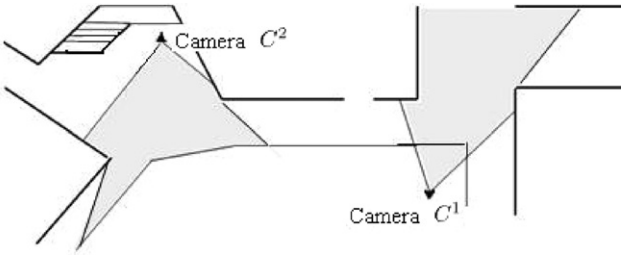

Fig. 14. Camera setup for scenario 3. It is an indoor/outdoor scenario. Camera C^1 was placed outdoor and camera C^2 was indoor.

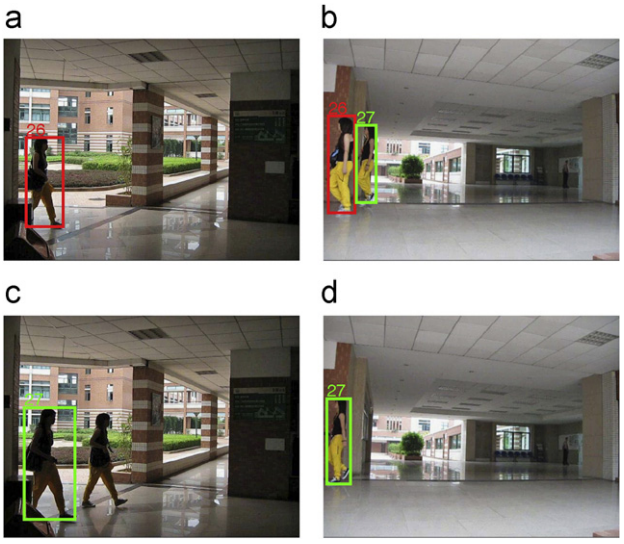

Fig. 15. Frames extracted from the testing phase of scenario 3. (a) and (c) show that camera C^2 views an area which is very dim, whereas (b) and (d) show that the scene viewed by camera C^1 is a hall where it is a covered area under shade. Therefore, the observed color information of the same person is completely different in the two views. Note that two people with the same color clothes from exiting camera C^1 to entering camera C^2 were assigned the correct label. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Matching accuracy.

	Total number (#)	Correct number (#)	Accuracy (%)
Scenario 1	36	32	88.89
Scenario 2	62	57	91.94
Scenario 3	39	34	87.18

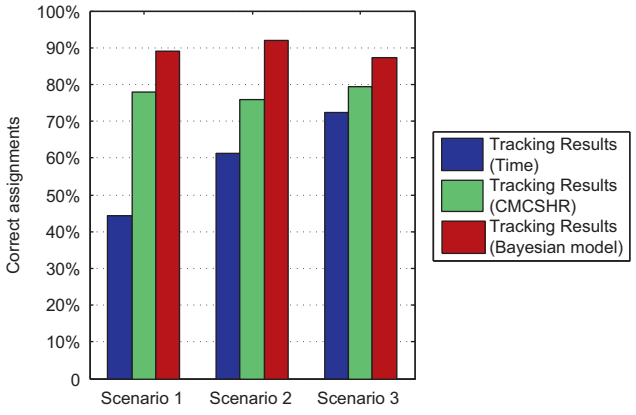

Fig. 16. Evaluation of the effects of different ingredients in our model. (1) Only using time model, (2) only using our proposed CMCSHR, and (3) using our proposed Bayesian model. The results show that the combination of spatio-temporal and appearance information using our Bayesian model is better than only using one of them for all scenarios.

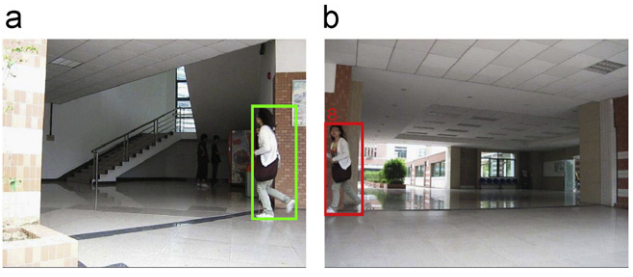

Fig. 17. An example of incorrect correspondence. It is because a person who ran across the two camera views.

an experiment that two people entered the entry FOV of camera C^2 almost simultaneously and one was occluded by the other. The experimental results at different occlusion percentages using the training and testing sequences in our three real scenarios are presented in Table 4. In this table, the correct correspondence is defined as the detected new object correctly corresponding to the hypothesis; that is the detected new object is exactly formed by the two people of this hypothesis, where the different occlusion

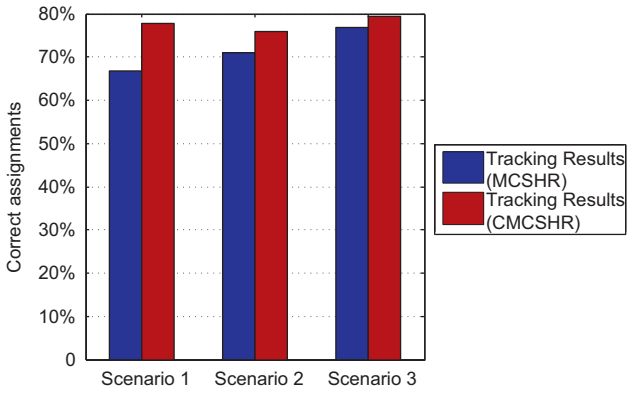

Fig. 18. Tracking accuracy: comparison between only using CMCSHR and only using MCSHR.

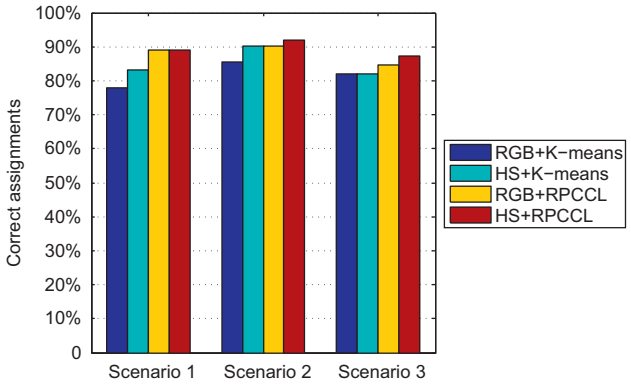

Fig. 19. Tracking accuracy: comparing RGB/HS+Kmeans with RGB/HS+RPCCL within our proposed Bayesian framework. The results show that CMCSHR (HS+RPPCL) performs the best.

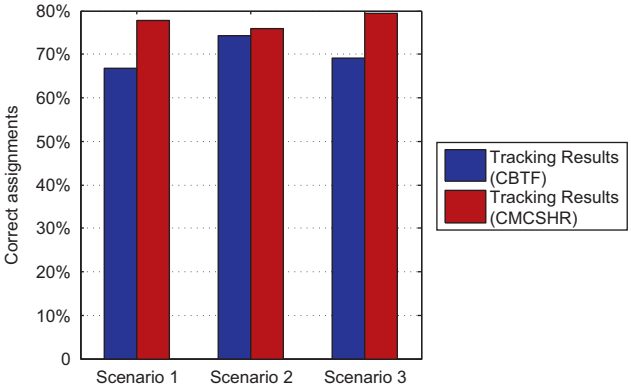

Fig. 20. A comparison of the tracking accuracy between the CMCSHR method and the CBTF approach.

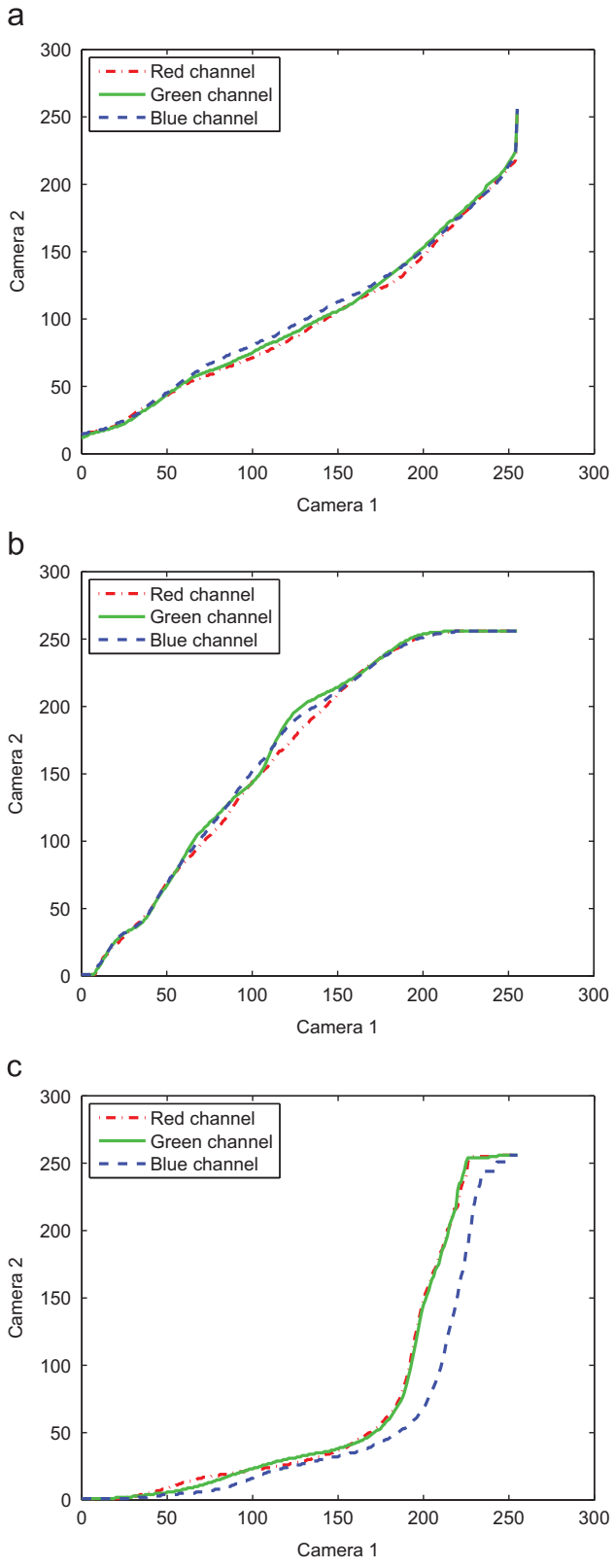

Fig. 21. (a) The transfer functions for three color channels from the exit FOV of camera C^1 to the entry FOV of camera C^2 in scenario 1. (b) The transfer functions in scenario 2. (c) The transfer functions in scenario 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

percentages were manually evaluated. Our results show that for different occlusion percentages our model achieves almost the same results. Fig. 23 shows some correct correspondence instances under occlusions. But if one person is completely occluded by another, the hypothesis formed by one person rather the two people has higher MAP. Fig. 24 shows an example of complete occlusion. In this figure, person 36 was completely occluded by person 35 in the entry FOV of camera C^2. The experimental result shows that the detected new object incorrectly corresponded to person 35, while the correct one should be the hypothesis formed by person 35 and person 36.

5. Conclusions and discussions

In this paper, we present a Bayesian model to solve the consistent labeling problem across multiple non-overlapping camera views. The Bayesian model unifies the spatio-temporal cue in terms of time of exit/entry, velocities of objects, and distance between entry field and exit field of two cameras and the visual appearance cue in terms of the proposed CMCSHR method. Our model neither requires each object is separated nor the trajectory of each object is estimated. An online algorithm for update of correspondence using OGM method is presented to perform the consistent labeling for the occlusion problem/group problem (i.e. the detected new object in the entry FOV of camera C^2 corresponds to a group hypothesis (more than one object)). Experiments on three different realistic scenarios validate the proposed approach and particularly show that the proposed formulation is able to deal with the consistent labeling against occlusion which occurs when a few people enter the entry FOV of camera C^2 almost simultaneously.

Although an online algorithm using OGM method is proposed to deal with the consistent labeling against occlusions, our proposed algorithm still has some limitations. Specifically, the matching between two objects in the algorithm is based on the similarity of their major colors. Hence, if O_{new} is a group object and its group hypothesis consists of people who wear completely the same uniform, the matching result between two objects after splitting will be uncertain. In this case, other features can be employed to solve this problem. Also, we should acknowledge that the success of our OGM based algorithm relies on the matching method such as CMCSHR, and hence if an incorrect matching is estimated, the tracking after an object split may be incorrect. Note that matching between two objects (particularly between two people) is still a largely unsolved problem in computer vision, and due to the use of spatio-temporal information, our OGM becomes applicable.

Currently, our system is still not specifically designed for a (extremely) busy surveillance scenario such as the hall of an airport. Note that in a crowded environment, objects are always seriously occluded each other and it is even hard to track a object using currently existing techniques in such a scenario. Also, in a busy surveillance scenario, the transition between cameras could be infinite and thus uncertain, and how to quantify the spatio-temporal cue in this case could be an open issue. Our future work would be on consistently labeling and tracking objects in busy surveillance scenarios across non-overlapping camera views.

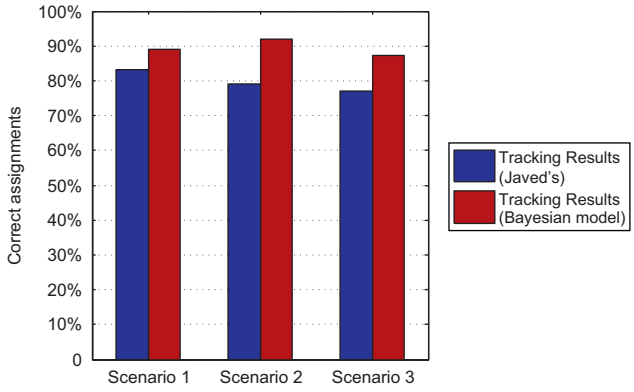

Fig. 22. Comparison of tracking accuracy for each of the three scenarios using our Bayesian model and Javed's approach in [22].

Table 4
Matching accuracy under different occlusion percentage.

	Total number	Correct number	Accuracy (%)
10% occluded	9	8	88.9
20% occluded	10	10	100
30% occluded	10	9	90
40% occluded	8	8	100
50% occluded	8	7	87.5
60% occluded	7	7	100
70% occluded	7	7	100
80% occluded	8	8	100
90% occluded	9	8	88.9
Completely occluded	1	0	0

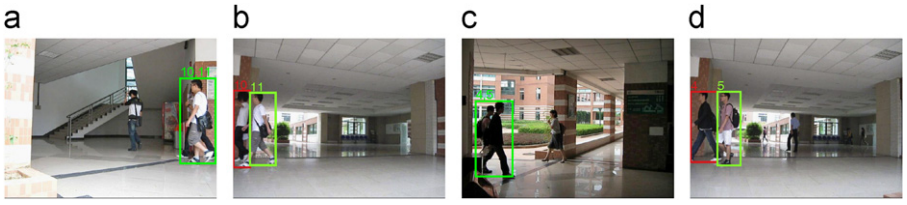

Fig. 23. Two instances of correct correspondence under occlusions. (a) and (b) show the correct correspondence under occlusions in scenario 2. (c) and (d) show the case in scenario 3.

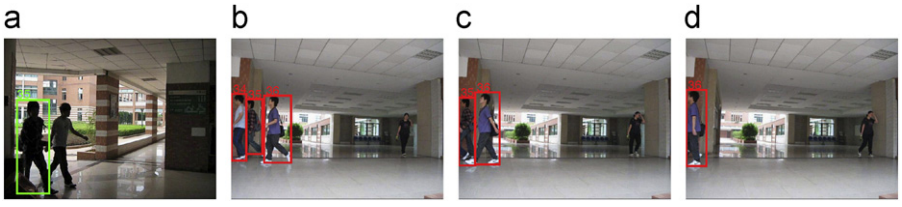

Fig. 24. An example of incorrect correspondence. It was caused by the new object (label 35) which was detected in the entry FOV of camera C^2 consisting of two people, that is, person 36 was completely occluded by person 35.

Acknowledgements

This project was supported by the NSFC-GuangDong (U0835005), NSFC (60803083), and GuangDong Program (2010B031000004). The authors would also like to thank the reviewers for their constructive advice.

Appendix A. Introduction of RPCCL

The basic idea of RPCCL [29] is that for each input, not only the winner of the seed points is updated to adapt the input, but also its nearest rival (i.e., the second winner) is penalized with the strength dynamically control by a mechanism. In this mechanism, the rival should be fully penalized if its distance to the winner is closer than the distance between the winner and the input; otherwise the penalization strength should be decreased as the rival distance to the winner increases. The algorithm using Euclidean distance is described as follows:

Step 1: Randomly take a sample $\mathbf{x}_t$ from the data set $D=\{\mathbf{x}_t\}_{t=1}^N$ and for $j=1,2,\ldots,k$, where k is the number of the seed points. Let

$$I(j|\mathbf{x}_t)=\begin{cases}1 & \text{if } j=c(\mathbf{x}_t),\\ -1 & \text{if } j=r(\mathbf{x}_t),\\ 0 & \text{otherwise,}\end{cases} \tag{30}$$

with

$$c(\mathbf{x}_t)=\underset{j}{\operatorname{argmin}}\gamma_j||\mathbf{x}_t-\mathbf{m}_j||^2,$$

$$r(\mathbf{x}_t)=\underset{j\neq c(\mathbf{x}_t)}{\operatorname{argmin}}\gamma_j||\mathbf{x}_t-\mathbf{m}_j||^2, \tag{31}$$

where $\gamma_j=n_j/\sum_{r=1}^k n_r$ is the relative winning frequency of the seed point $\mathbf{m}_j$ in the past, and n_j is the cumulative number of the occurrences of $I(j|\mathbf{x}_t)=1$ in the past.

Step 2: Update the winner $\mathbf{m}_c(i.e., I(c|\mathbf{x}_t)=1)$ and its rival $\mathbf{m}_r$ only by

$$\mathbf{m}_u^{new}=\mathbf{m}_u^{old}+\Delta\mathbf{m}_u,\quad u=c,r, \tag{32}$$

with

$$\Delta\mathbf{m}_c=\alpha_c(\mathbf{x}_t-\mathbf{m}_c), \tag{33}$$

$$\Delta\mathbf{m}_r=-\alpha_c p_r(\mathbf{x}_t)(\mathbf{x}_t-\mathbf{m}_r), \tag{34}$$

where α_c is the learning rate. These two steps are repeated for each input until $I(j|\mathbf{x}_t)$ is converged. The measurement of the rival penalization strength is as follows:

$$p_r(\mathbf{x}_t)=\frac{\min(|\mathbf{m}_c-\mathbf{m}_r|,|\mathbf{m}_c-\mathbf{x}_t|)}{|\mathbf{m}_c-\mathbf{m}_r|}. \tag{35}$$

The RPCCL penalizes the rivals such that extra seed points are automatically driven far away from the input data set.

References

[1] O. Javed, K. Shafique, M. Shah, Appearance modeling for tracking in multiple non-overlapping cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 26–33.
[2] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, S. Maybank, Principal axis-based correspondence between multiple cameras for people tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (4) (2006) 663–670.
[3] Y. Jo, J. Han, W. Nam, Object handoff between uncalibrated views without planar ground assumption, Pattern Recognition Letters 29 (2008) 2099–2108.
[4] S.M. Khan, M. Shah, Consistent labeling of tracked objects in multiple cameras with overlapping fields of view, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (10) (2003) 1355–1360.
[5] S. Calderara, R. Cucchiara, A. Prati, Bayesian-competitive consistent labeling for people surveillance, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2) (2008) 354–360.
[6] R. Munoz-Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, A. Carmona-Poyato, Particle filtering with multiple and heterogeneous cameras, Pattern Recognition, 43 (7) (2010) 2390–2405.
[7] A. Mittal, L.S. Davis, M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo, in: Proceedings of European Conference on Computer Vision, 2002, pp. 18–36.
[8] K. Nummiaro, E. Koller-Meier, T. Svoboda, D. Roth, L. Van Gool, Color-based object tracking in multi-camera environments, in: Proceedings of the 25th DAGM Symposium on Pattern Recognition, 2003, pp. 591–599.
[9] D. Makris, T. Ellis, J. Black, Bridging the gaps between cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 205–210.
[10] X. Wang, K. Tieu, W.E.L. Grimson, Correspondence-free activity analysis and scene modeling in multiple camera views, IEEE Transactions on Pattern Analysis and Machine Intelligence 1 (1) (2009) 1–17.
[11] A. Rahimi, T. Darrell, Simultaneous calibration and tracking with a network of non-overlapping sensors, in: IEEE Conference on Computer Vision and Pattern Recognition, 2004, pp. 187–194.
[12] V. Kettnaker, R. Zabih, Bayesian multi-camera surveillance, in: IEEE Conference on Computer Vision and Pattern Recognition, 1999, pp. 252–259.
[13] C. Madden, E.D. Cheng, M. Piccardi, Tracking people across disjoint camera views by an illumination-tolerant appearance representation, Machine Vision and Applications 18 (2007) 233–247.
[14] E.D. Cheng, M. Piccardi, Disjoint track matching based on a major color spectrum histogram representation, Optical Engineering 46 (4) .
[15] B. Prosser, S. Gong, T. Xiang, Multi-camera matching using bi-directional cumulative brightness transfer functions, in: BMVC, 2008.
[16] K. Jeong, C. Jaynes, Object matching in disjoint cameras using a color transfer approach, Machine Vision and Application 19 (2008) 443–455.
[17] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, P. Tu, Shape and appearance context modeling, in: International Conference on Computer Vision, 2007, pp. 1–8.
[18] N. Gheissari, T. Sebastian, R. Hartley, Person reidentification using spatio-temporal appearance, Computer Vision and Pattern Recognition (2006) 1528–1535.
[19] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision, 2008, pp. 262–275.
[20] W.-S. Zheng, S. Gong, T. Xiang, Associating groups of people, in: British Machine Vision Conference, 2009.
[21] A. Gilbert, R. Bowden, Tracking objects across cameras by incrementally learning inter-camera colour calibration and patterns of activity, in: European Conference on Computer Vision, 2006, pp. 125–136.
[22] O. Javed, K. Shafique, Z. Rasheed, M. Shah, Modeling inter-camera space-time and appearance relationships for tracking across non-overlapping views, Computer Vision and Image Understanding 109 (2008) 146–162.
[23] K. Chen, C. Lai, Y. Hung, C. Chen, An adaptive learning method for target tracking across multiple cameras, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.
[24] S. Calderara, A. Prati, R. Cucchiara, HECOL: homography and epipolar-based consistent labeling for outdoor park surveillance, Computer Vision and Image Understanding 111 (2008) 21–42.
[25] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., Wiley-Interscience, 2000.
[26] M.J. Swain, D.H. Ballard, Indexing via color histograms, in: Proceedings of the International Conference on Computer Vision, 1990, pp. 390–393.
[27] R.C. Gonzalez, R.E. Woods, Digital Image Processing, second ed., Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1992.
[28] E.W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretability of classifications, Biometrics 21 (1965) 768–780.
[29] Y.M. Cheung, Rival penalization controlled competitive learning for data clustering with unknown cluster number, in: Proceedings of 9th International Conference on Neural Information Processing, vol. 2, 2002, pp. 467–471.
[30] G. Tzortzis, A. Likas, The global kernel k-means clustering algorithm, in: Proceedings of the IEEE International Joint Conference on Neural Networks, 2008, pp. 1977–1984.
[31] Y. Zhang, Z. Liu, Self-splitting competitive learning: a new on-line clustering paradigm, IEEE Transactions on Neural Networks 13 (2) (2002) 369–380.
[32] A. Likas, N. Vlassis, J.J. Verbeek, The global k-means clustering algorithm, Pattern Recognition 36 (2) (2003) 451–461.
[33] C. Hue, J.L. Cadre, P. Prez, Sequential monte carlo methods for multiple target tracking and data fusion, IEEE Transactions on Signal Processing 50 (2) (2002) 309–325.
[34] X.J. Wan, Y.X. Peng, A new retrieval model based on TextTiling for document similarity search, Journal of Computer Science & Technology 20 (4) (2005) 552–558.
[35] L. Lovasz, M.D. Plummer, Matching Theory, North-Holland, Amsterdam, 1986.
[36] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 747–757.

Guoyun Lian received his M.Sc. in computer science from Kunming University of Science and Technology in 2008. He is currently a Ph.D. candidate in information science and technology in Sun Yat-sen University, Guangzhou, China. His current research interests include computer vision, pattern recognition, digital image processing and analysis, visual surveillance, etc.

Jianhuang Lai received his M.Sc. degree in applied mathematics in 1989 and his Ph.D. in mathematics in 1999 from Sun Yat-sen University, China. He joined Sun Yat-sen University in 1989 as an Assistant Professor, where currently, he is a Professor with the Department of Automation of School of Information Science and Technology and vice dean of School of Information Science and Technology. He had successfully organized the International Conference on Advances in Biometric Personal Authentication'2004, which was also the Fifth Chinese Conference on Biometric Recognition (Sinobiometrics'04), Guangzhou, in December 2004. He has taken charge of more than five research projects, including NSF-Guangdong (no. U0835005), NSFC (nos. 60144001, 60373082, 60675016), the Key (Keygrant) Project of Chinese Ministry of Education (no. 105134), and NSF of Guangdong, China (nos. 021766, 06023194). He has published over 80 scientific papers in the international journals and conferences on image processing and pattern recognition. His current research interests are in the areas of digital image processing, pattern recognition, multimedia communication, wavelet and its applications. He serves as a standing member of the Image and Graphics Association of China and also serves as a standing director of the Image and Graphics Association of Guangdong.

Wei-Shi Zheng is a Postdoctoral Researcher at the Department of Computer Science, Queen Mary University of London, UK. He is now working on the European SAMURAI Research Project with Prof. Shaogang Gong and Dr. Tao Xiang. Prior to that, he worked on subspace methods for face recognition with supervision by Prof. JianHuang Lai and received his Ph.D. degree in Applied Mathematics at Sun Yat-Sen University, China, 2008. He has been a visiting student working with Prof. Stan Z. Li at the Institute of Automation, Chinese Academy of Sciences, and an exchanged research student working with Prof. Pong C. Yuen at Hong Kong Baptist University. He was awarded the HP Chinese Excellent Student Scholarship 2008. His current research interests are in object association and categorization for visual surveillance. He is also interested in discriminant/sparse feature extraction, dimension reduction, kernel methods in machine learning, transfer learning, and face image analysis.