Multimed Tools Appl (2017) 76:217–241
DOI 10.1007/s11042-015-3015-5

Online learning of dynamic multi-view gallery for person Re-identification

Yanna Zhao[1,2] · **Xu Zhao**[2] · **Zongjie Xiang**[2] · **Yuncai Liu**[2]

Received: 13 January 2015 / Revised: 4 September 2015 / Accepted: 19 October 2015 /
Published online: 6 November 2015

Abstract Person re-identification receives increasing attentions in computer vision due to its potential applications in video surveillance. In order to alleviate wrong matches caused by misalignment or missing features among cameras, we propose to learn a multi-view gallery of frequently appearing objects in a relatively closed environment. The gallery contains appearance models of these objects from different cameras and viewpoints. The strength of the learned appearance models lies in that they are invariant to viewpoint and illumination changes. To automatically estimate the number of frequently appearing objects in the environment and update their appearance models online, we propose a dynamic gallery learning algorithm. We specifically build up two datasets to validate the effectiveness of our approach in realistic scenarios. Comparisons with benchmark methods demonstrate promising performance in accuracy and efficiency of re-identification.

Keywords Person re-identification · Dynamic gallery learning · Feature correspondence · Clustering

✉ Yanna Zhao
yannazhao@outlook.com

✉ Xu Zhao
zhaoxu@sjtu.edu.cn

Zongjie Xiang
zongzong_1984_111@hotmail.com

Yuncai Liu
whomliu@sjtu.edu.cn

[1] School of Information Science and Engineering, Shandong Normal University, Jinan, China

[2] School of Automation, Shanghai Jiao Tong University, Shanghai, China

1 Introduction

Large scale video surveillance has been augmented by the eager request for security purposes. A critical issue of this task is to automatically associate objects across disjoint camera views, known as person re-identification. Typically, person re-identification is the problem of discovering a target person (a probe) from a crowd of people (the gallery) captured by different cameras, at different locations based on their appearance similarity. This problem is of great interest to numerous computer vision applications, including public security, long-term object tracking, threat detection and behavior analysis. Although person re-identification has been studied for many years, it remains a challenging and unsolved problem due to visual ambiguities and uncertainties caused by viewpoint and pose variations, illumination changes and occlusions as shown in Fig. 1.

The main difficulty of person re-identification lies in the severe variations from different cameras and viewpoints that can cause significant changes in appearance. Directly matching the features of object images from different cameras is unreliable due to feature misalignment or even missing features. For example, in Fig. 1, the central region of image (a8) is gray in camera view A, while it becomes plaid shirt in image (b8) in camera view B. Recent studies [1, 3, 5, 7, 9, 11, 19, 24, 25, 31, 34] handle the problem of cross-view variations by seeking a more distinct and reliable low-level representation of human appearance. The most commonly used features include: color [5, 7, 9, 19], texture [5, 9, 11, 25], covariance features [1, 19, 25], HOG (Histogram of Oriented Histograms) like signatures and interest points [11, 34]. However, it is extremely difficult to compute both distinct and reliable low-level features under severe variations in different camera views.

To bridge the human appearance changes across cameras, we present an effective method for person re-identification. Our method focuses on re-identification in a relatively closed environment, such as a school, an office building and a house, where most observed objects appear repeatedly in different locations. We refer to these objects as *regular persons* and their features as *regular features*. Even in a relatively closed environment, unknown or new objects could also appear. We refer to these objects as *strangers* and their features are referred to as

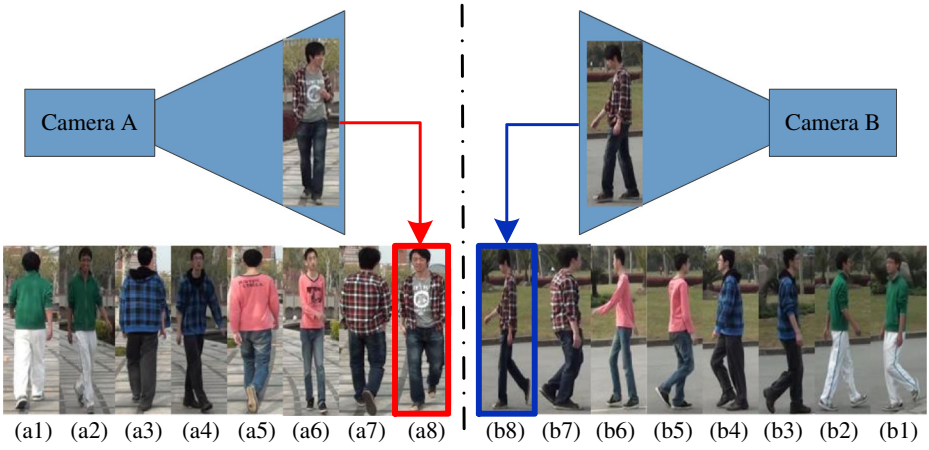

Fig. 1 Person re-identification using our multi-view gallery. Images on the left of the dashed black line are from camera view A and those on the right are from camera view B. Camera A captures the front and back views of an object and camera B captures the left and right views. Our method finds the identity of the target by comparing visual similarity with the appearance feature in the corresponding camera and viewpoint

stranger features. We assume that a camera network is mounted in this environment. These cameras will capture different aspects of appearance for those regular persons as they move from one Field of View (FoV) to another. With this appearance information, we can build a multi-view gallery for these regular persons. This gallery contains the appearance models of these objects in each camera and viewpoint. When a target enters, we can confirm his/her identity by searching the gallery in the corresponding camera and viewpoint. For example, in Fig. 1, camera A records the frontal and back views of the objects while camera B records the left and right views. The appearance models corresponding to the two cameras and four views are stored in our multi-view gallery. The target person in camera A, with a frontal view, finds his identity by comparing appearance models in the same camera and viewpoint. If the target object cannot be associated with any one of the gallery models, we will treat the object as a stranger. In some cases, like threat detection, suspicious check and behavior analysis, keep tracking of the strangers over different cameras is also important as they are more suspicious. Using the appearance information of the object and the topological information of the camera network, keep tracking of the strangers is also achievable.

The number of regular persons and the appearance of the regular persons will change over time, e.g., the illumination conditions are significantly different for images (a1) and (a2) in Fig. 1. In order to deal with these changes, we propose a dynamic gallery learning algorithm. The algorithm iterates between the following two steps: (1) Automatically estimate the current number of regular persons by employing clustering quality analysis. When a regular person leaves or a new object comes into the environment for a long time, the corresponding appearance models will be deleted or added, yielding a new gallery. (2) Establish the correspondence between the newly obtained gallery and the old gallery. The newly obtained gallery also incorporates illumination changes of each camera view over time. Experimental results validate the effectiveness of the algorithm.

This paper extends our previous work [33], which focused on generating a static multi-view gallery for re-identification. In static gallery learning, we need to set the number of regular persons manually, which prevents the algorithm from working automatically. Besides, the appearance models contained in the static gallery cannot be updated along time. In this paper, we propose a new algorithm to learn and update the gallery dynamically. In addition, more comparisons are made with existing methods. Our approach is inherently different from existing works that focus on developing reliable features to deal with cross view variations. Since the appearance models for each object are specific to a certain camera and viewpoint, existing feature representations can be adopted to describe an object. In summary, the main contributions of this work are:

- While most existing works on person re-identification focus on new visual features and models that deal with cross-view variations, we propose to build a multi-view gallery by accumulating appearance information from different cameras and viewpoints over time and use this gallery to perform re-identification. We show that the multi-view gallery improves person re-identification accuracy because of its ability to deal with unreliable matching caused by feature misalignment or missing features.
- We propose an algorithm to build and update the gallery dynamically. The dynamic gallery is adaptive to changes of appearance and the number of regular persons in the environment. To the best of our knowledge, this is the first attempt to learn a dynamic multi-view gallery for person re-identification. Experiments on video sequences collected from both indoor and outdoor environments demonstrated the effectiveness of the algorithm.

The rest of the paper is organized as follows. Section 2 describes the related work on person re-identification. Section 3 presents the idea of gallery learning and the dynamic gallery learning algorithm, which is followed by experiments in Section 4. Finally, conclusions and further perspectives are given in Section 5.

2 Related work

Person re-identification researches can be categorized into two classes: learning-based methods and direct methods. The first class uses a set of training images captured from different objects to learn a discriminative feature space. Essentially, these techniques assume that knowledge extracted from the training set could be generalized to new samples. Discriminative models like boosting and SVM are widely used for feature learning [2, 4, 12, 14, 20, 28, 29]. Gray et al. and Bak et al. [2, 4, 12] employed boosting strategy to select the most discriminative subset of features. Prosser et al. [28] formulated person re-identification as a ranking problem. An informative subspace was learned where the potential true match gets highest ranking. Metric learning algorithms have also been applied to learn task-specific distance functions to suppress cross-view variations [8, 18, 21, 22, 27, 35]. Zheng et al. [35] proposed the Probabilistic Relative Distance Comparison (PRDC) that maximizes the probability of a pair of true match having a smaller distance than that of a wrong matched pair. Li et al. [22] presented to learn a specific metric for each query-candidate set by transferred metric learning framework. Mignon et al. [27] designed a metric for identification under pairwise constraints in high dimensional space. Li et al. [21] proposed to learn a decision function for verification combining distance metric learning and locally adaptive thresholding rule.

One problem with these methods is that, they require training data with identity labels, which is impractical in large scale surveillance scenarios. Besides, inter-camera variations can be significantly different, knowledge learned from the training data cannot be generalized to testing samples. As a result, they have to be frequently re-trained/updated when facing realistic scenarios.

The second class of methods is direct methods, which focus on reliable feature representation [1, 3, 5, 7, 9, 11, 19, 24, 25, 31, 34]. These methods do not consider training data but rather working on each object directly. Wang et al. [31] extracted discriminative features by modeling spatial distributions of appearance relative to body parts. Farenzena et al. [9] proposed Symmetry Driven Accumulation of Local Features (SDALF). They partitioned object images into symmetry and asymmetry parts. Color and texture were combined to describe each part, yielding state-of-the-art results on several widely used datasets. Cheng et al. [7] estimated human body configurations utilizing the pictorial structures and computed similar visual features as SDALF on different body parts. Bak et al. [1] combined local statistics of color and gradient to construct a covariance descriptor. Ma et al. [25] proposed to combine biologically inspired features and covariance descriptors to handle background and illumination changes. Zhao et al. [34] developed an unsupervised saliency learning strategy to extract discriminative features. These works share the same general idea: a feature vector to represent an object and a distance measure to compare similarity. When selecting the feature vector one has to make sure that it is both discriminative and invariant.

In general, learning-based methods produce higher performances than direct methods. However, they are limited to practical usage. The main limitation of the learning-based methods is that image representations they employed are complex, which prohibit them from

real-time applications. To cope with these problems, we propose an effective and efficient framework for person re-identification and apply it to video sequences acquired when people performing their daily activities.

Our method follows the classical scheme of Detection-Recognition-Identification (DRI). There are other methods that follow the DRI scheme for robust matching and tracking [10, 13, 15, 16]. For instance, Javed et al. [15] explicitly modeled the brightness transfer function for a pair of cameras to compensate for illumination variations. Similarly, Jeong et al. [16] treated the appearance model distortion between two non-overlapping cameras by learning a color transfer function. Both methods assume that the conditions under which these functions are estimated remain fixed, which is inappropriate in real surveillance. Hamdoun et al. [13] performed person re-identification by matching SURF interest-points descriptors collected on short video sequences. The method proposed by Gandhi et al. [10] shares a similar spirit to our work, which aims to build a signature for each person by combining information of the object from different camera views. But their panoramic appearance maps are built using multiple overlapping cameras. The panoramic appearance maps are not available when the cameras are of non-overlapping views.

3 Dynamic multi-view gallery learning for person Re-identification

Once an object has been detected and tracked with a bounding box in multiple frames, we begin the process to learn a dynamic multi-view gallery. The building process mainly consists three phases:

(1) Collect appearance features of each object using the tracking results and separate the features into four viewpoints.
(2) Learn a multi-view gallery that contains the appearance information of each object from different cameras and viewpoints.
(3) Dynamically update the multi-view gallery in order to adapt to the changes over the number of regular persons and appearance of the regular persons.

In the following, we first illustrate our problem setup and feature collection. Then we describe and analyze the above mentioned three phases accordingly.

3.1 Problem setup and feature collection

Our multi-view gallery divides appearance features into four viewpoints: left, right, front and back. The features are collected from the tracks of each object. From each track, we can extract at most 4 average appearance features: A_{left}, A_{right}, A_{front} and A_{back}, which describe an object's appearance information from 4 different viewpoints. If the object does not change his/her walking direction within the FoV of a camera, we can extract only one average feature from this track. By "average", we mean that the features taken from all the frames of this track under the same viewpoint are averaged.

We use the Fuzzy Space Color Histogram (FSCH) [32] to describe an object in each camera and viewpoint because of its efficiency and effectiveness. FSCH feature A can be computed by using the following functions:

$$A(R_{bc}, G_{bc}, B_{bc}, x_{bc}, y_{bc}) = \sum_{R,G,B,x,y} h(R,G,B,x,y) w_1(R-R_{bc}) w_2(G-G_{bc}) w_3(B-B_{bc}) w_4(x-x_{bc}) w_5(y-y_{bc}) \quad (1)$$

$$h(R,G,B,x,y) = \begin{cases} 1/ToatalPixels, pixel(x,y) = [R,G,B] \\ 0 \qquad\qquad , otherwise \end{cases} \quad (2)$$

$$\begin{cases} w_i(x) = w_i(-x) \\ w_i(0) = 1 \\ w_i(x_1) \le w_i(x_2), |x_1| > |x_2| \end{cases} \quad (3)$$

where *R, G, B* are RGB values of the pixel at position (*x, y*), the subscript *bc* denotes the bin center, function *h* is the space color histogram, *TotalPixels* is the total number of pixels in the image, and $w_i(x)$ is the unary membership function. The FSCH feature incorporates space information into the 3D color histogram and is characterized by "soft quantization".

The FSCH feature is stored as a structure *feature* defined as:

$$feature = (A, C, V, T_{EN}, T_{EX}, L_{EN}, L_{EX}, ID_P) \quad (4)$$

where *A* is the appearance feature, *C* and *V* are the camera and viewpoint identities, T_{EN}/T_{EX} is the entrance/exit timestamp (the timestamp when the corresponding track begins/ends), L_{EN}/L_{EX} is the entrance/exit point (the start/end point of the corresponding track), and ID_P is the object's identity.

The features are stored in the *feature pool*. After collecting a set of *feature*s, we separate them into several *group*s so that each member within the same *group* shares the same camera identity *C* and viewpoint identity *V*, as defined in Eq. (5).

$$group(C_i, V_j) = \{feature | feature.C = C_i, feature.V = V_j\} \quad (5)$$

3.2 Multi-view gallery learning

The aim of this phase is to establish a multi-view gallery that contains the appearance information of the objects in different cameras and viewpoints using the features collected in section 3.1. First, we cluster the member features in each *group* according to appearance similarity. Second, we associate the clusters of the same object in different *group*s using the topology information. Finally, we merge all the clusters so that the final number of clusters equals the number of regular persons.

Feature clustering according to appearance The member features in each *group* are separated into several clusters by calculating their appearance similarity using the following function:

$$Similarity(A_1, A_2) = \frac{\sum_{i=1}^{N} \left(A_1^i - \overline{A}_1\right)\left(A_2^i - \overline{A}_2\right)}{\sqrt{\sum_{i=1}^{N} \left(A_1^i - \overline{A}_1\right)^2 \sum_{i=1}^{N} \left(A_2^i - \overline{A}_2\right)^2}} \quad (6)$$

where $\overline{A} = \frac{1}{N}\sum_{i=1}^{N} A^i$ is the average feature, N is the length of the feature vector, and i indicates the i-th dimension of the feature vector. In Fig. 2, two cameras C_1, C_2 and four viewpoints V_1, V_2, V_3, V_4 yield eight *group*s. The symbols ○, △ and ☆ denote three clusters. Each cluster represents the appearance of a regular person in the corresponding camera C_i and viewpoint V_j. The number of clusters equals the number of regular persons. We manually set the number of clusters in advance.

Associating the clusters in different groups As illustrated in Fig. 2, different objects within each *group* have been distinguished. The clusters of the same object among different *group*s must be associated. This step is performed using the topology information of the camera network.

For example, cluster △ in *group* (C_1,V_2) and cluster △ in *group* (C_2,V_1) are associated (see $linkage_1$ in Fig. 2). Figure 3 explains the reason for this association. The small circles in cluster △ represent the features contained therein. These features, which are members of the clusters, shall be referred to as member features. There is a passage between camera C_1 and camera C_2, which requires an average of 10 s to walk through. Four pairs of member features in cluster △ in different cameras are connected by directed arcs, because the actual time interval between T_{EX} and T_{EN} roughly matches the average time spent in the passage. This phenomenon occurs four times and only five member features exist in cluster △ of *group* (C_2,V_1), indicating a strong correspondence between the two clusters. Therefore, we can infer that cluster △ in *group* (C_1,V_2) and cluster △ in *group* (C_2,V_1) belong to the same object and can be associated.

Specifically, assume that F and G are two clusters in two different groups:

$$\begin{cases} F \subset group_1 \\ G \subset group_2 \end{cases} \quad group_1 \neq group_2 \tag{7}$$

f and g are two *features* in F and G respectively. Let $T_{average}$ denotes the average time spent from the exit point of f to the entrance point of g. If the passage does not exist, $T_{average}$ is infinity. As f and g are stored as structure, the actual exit and entrance timestamps are $f.\ T_{EX}$ and $g.\ T_{EN}$ respectively (see Eq. 4). So the difference between the actual time gap and the average time can be computed as:

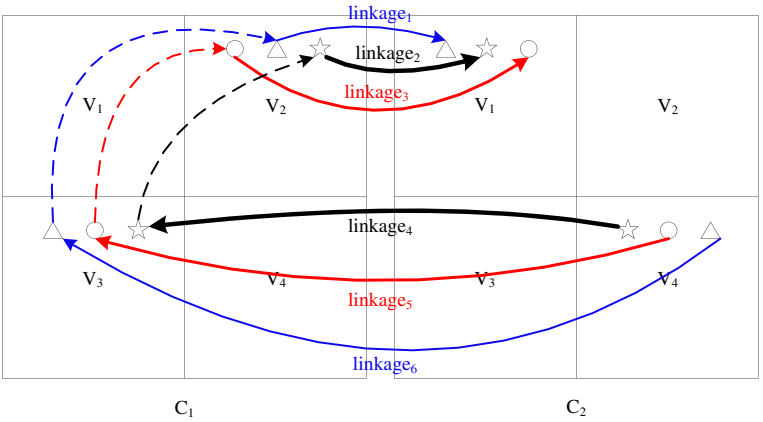

Fig. 2 Illustration of multi-view gallery learning

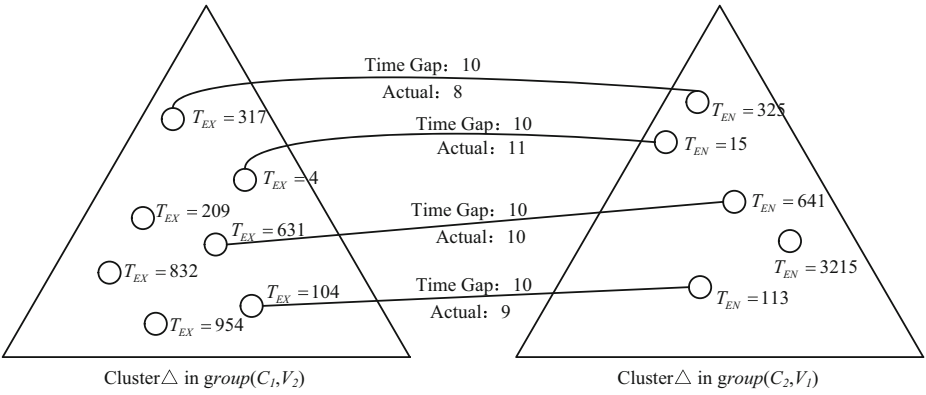

Fig. 3 Associate clusters in different groups

$$T_d = \left| g.T_{EN} - f.T_{EX} - T_{average} \right| \tag{8}$$

The correspondence strength of f and g is computed as:

$$str(f,g) = \begin{cases} 1 - T_d / \sigma T_{average}, & T_d / T_{average} \le \sigma \\ 0, & otherwise \end{cases} \tag{9}$$

where the parameter σ is a constant ranging from 0 to 1, which is used to control the tolerance to T_d. We set it to 0.4 in the experiments. A small T_d will lead to large *str*. The correspondence strength of clusters F and G is computed as:

$$STR(F,G) = \frac{\sum_{m=1,n=1}^{M,N} str(f_m, g_n)}{\min(M,N)} \tag{10}$$

where M and N are the numbers of *features* in F and G respectively. If *STR (F,G)* is larger than a threshold (set to 0.3 in our experiment), we can conclude that cluster F and cluster G belong to the same object. As shown in Fig. 2, the clusters belong to the same object are connected by solid directed arcs.

Merging into the final clusters It is possible that the number of linkages is higher than the number of regular persons after feature correspondence. As shown in Fig. 2, the 6 linkages should be further merged into 3 because there are three regular persons in the example. If the number of regular persons is N_P, we choose N_P linkages as the core for the upcoming merging process.

Given N_P core linkages, the remaining is merged with the most similar one according to the average appearance similarity using Eq. (6). As depicted by the three dashed directed arcs in Fig. 2, $linkage_1$ and $linkage_6$, $linkage_2$ and $linkage_4$, and $linkage_3$ and $linkage_5$ are merged into three final linkages.

The output of the framework is a *gallery* of regular persons defined as:

$$gallery = \left\{ A_{P_i,C,V} | i \in [1, N_P], C \in [1, N_C], V \in [1, N_V] \right\} \tag{11}$$

where N_P is the number of regular persons, N_C is the number of cameras, and N_V is the number of viewpoints. Given the person identity ID_P, camera identity C and viewpoint identity V, we

can find the appearance model A of the person from the *gallery*. The appearance models are the features averaged over all member features in each cluster. Figure 4 gives an overall review of establishing the multi-view gallery. The procedures of static gallery learning are summarized in Algorithm 1.

Algorithm 1: The procedures of the static gallery learning

Input:

Appearance feature A, camera/viewpoint identity: C, V; entrance/exit timestamp: T_{EN}, T_{EX}; entrance/exit point: L_{EN}, L_{EX}; person identity: ID_P.

Output:

A multi-view gallery of regular persons

1: Separate features in each camera and viewpoint into several clusters according to appearance feature A. The number of clusters equals the number of regular persons.

2: Associate two clusters in different cameras and viewpoints using the topology information of the camera network. Each association is denoted by a *linkage*.

3: Merge all the *linkage*s into N_P core linkages according to the average appearance similarity of the member features contained therein.

4: Establish the multi-view *gallery* of regular persons, which contains the appearance features of the regular persons in the corresponding camera C and viewpoint V.

3.3 Dynamic learning of the multi-view gallery

For the multi-view gallery learning in section 3.2, we need to set the number of regular persons manually. Human intervention prevents the whole algorithm from working automatically. In addition, gallery learning is done during the training phase and could not be updated along time.

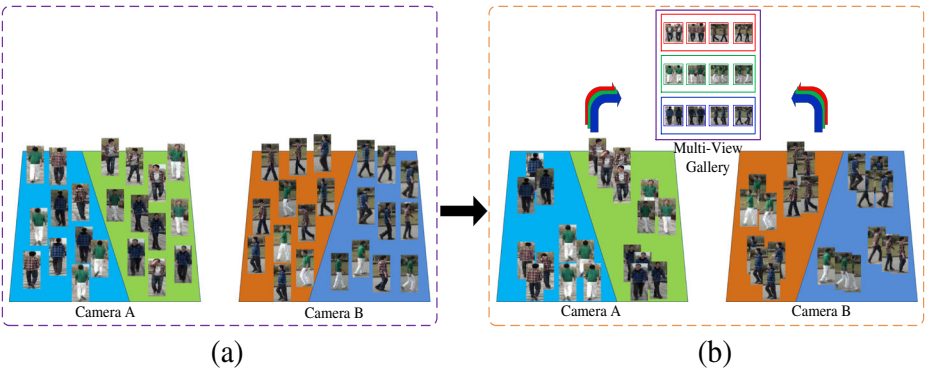

Fig. 4 Establishing the multi-view gallery. **a** Collecting image features from different cameras and viewpoints. **b** Image clustering according to visual similarity, and establishing the multi-view gallery

In this section, we present a dynamic algorithm to learn the multi-view gallery of regular persons. The algorithm iterates the following two phases. (1) Estimating the number of regular persons: given the *feature pool* and the initial interval $[n_1, n_2]$, automatically decide the best number of clusters n_{best}. (2) Establishing the correspondence between the new gallery and the old gallery: given $OldGallery = \{P_1, P_2, \cdots, P_{N_P}\}$ and $NewGallery = \left\{P_1{}', P_2{}', \cdots, P_{N_P}{}'\right\}$, find the *correspondence* among members in each gallery.

Estimating the number of regular persons Given the *feature pool*, we separate its member features into several groups so that each member within the same group shares the same camera identity C and viewpoint identity V. In each *group* (C, V), if we cluster its features according to appearance A, the number of clusters should be equal to the number of regular persons in that group. Therefore, the problem can be modeled as: how many clusters exist in a given set of data. For this purpose, we use k-means to estimate the number of clusters and assume that the true number of clusters n_{true} lies within an interval of positive integers, denoted as $[n_1, n_2]$. n_{true} is estimated through the interval $[n_1, n_2]$ one by one. In this phase, we first discuss how to set the interval dynamically in order to contain the actual number of regular persons, and then we discuss how to make the best decision of n_{true}.

Our dynamic gallery learning algorithm is data triggered. Assume that the r-th estimated number of regular persons is $n_{best}\ (r)$. When a new track is recorded, the *feature pool* will be updated and the gallery will be learned once again. The new n_1 and n_2 will be set according to:

$$\begin{aligned} n_1(r+1) &= n_{best}(r) - n_{half} \\ n_2(r+1) &= n_{best}(r) + n_{half} \end{aligned} \tag{12}$$

where n_{half} is a constant larger than 1, which controls the width of the interval.

When a new track is added into the *feature pool*, the true number of regular persons $n_{true}\ (r+1)$ does not deviate significantly from the previous true number $n_{true}\ (r)$ (see Eq. (13)). It may remain unchanged because there is no new regular person appears, increase by 1 because a new regular person is identified, or decrease by 1 because an existing regular person has not appeared for a long time and his features have been overwritten in the *feature pool* by the features of the newly appearing person.

$$|n_{true}(r+1) - n_{true}(r)| \leq 1 \tag{13}$$

If the previous estimation of the number of regular persons is correct, from Eqs. (12) and (13) we can infer that the interval $[n_1, n_2]$ will include the true number of regular persons, as shown in Eq. (14):

$$n_{true}(r) = n_{best}(r) \Rightarrow n_{true}(r+1) \in [n_1(r+1), n_2(r+1)] \tag{14}$$

Otherwise, if $n_{best}\ (r)$ is incorrectly estimated and deviates too much from $n_{true}\ (r)$, then $n_{true}\ (r+1)$ may lie outside the interval $[n_1\ (r+1),\ n_2\ (r+1)]$. This inevitably happens at the initialization of n_{best}, when we manually set n_1 and n_2. At initialization, we set $n_{best}\ (0)=2$, because k-means clustering requires the minimum number of clusters to be set to 2. With more tracks being recorded, the interval $[n_1, n_2]$ would slowly move towards n_{true} and finally include it. Because the selected number from $[n_1, n_2]$ closest to n_{true} will have a higher evaluation score, which will be discussed later.

After getting the interval $[n_1, n_2]$, we define a *Score* function to evaluate the quality of the clustering results [17] in one *group (C, V),* and n_{true} equals the best number of clusters $n_{best,}$ which gets the highest evaluation score according to:

$$n_{best} = \underset{n_1 \le n_k \le n_2}{\operatorname{argmax}} \left(Score(results(data, n_k)) \right) \tag{15}$$

where *results* is the *k*-means clustering result of the features *data* with n_k clusters. Assume that the total number of members in *data* is N_{data}, and d_i is assigned to cluster *k*. Then *Score* is the average clustering quality Q of each member d_i in *data*:

$$Score(results) = \sum_{i=1}^{N_{data}} Q(results, d_i)/N_{data} \tag{16}$$

$$Q(results, d_i) = (dist_{inter} - dist_{intra})/\max(dist_{intra}, dist_{inter}) \tag{17}$$

$$dist_{inter}(results, d_i) = \min_{j \neq k} \left(distance\left(center_j, d_i\right) \right) \tag{18}$$

$$dist_{intra}(results, d_i) = \sum_{d \in k, d \neq d_i} distance(d, d_i)/(N_k - 1) \tag{19}$$

where $dist_{inter}$ is the inter-cluster distance between d_i and the *center* of its nearest neighbor cluster (excluding cluster *k*), $dist_{intra}$ is the average distance between d_i and other members in the same cluster *k*, and N_k is the number of data in cluster *k*. The average clustering quality Q of each member d_i is a measure of how similar that member to members in its own cluster vs. members in other clusters. In practice, we use the Matlab built in function "silhouette" to get this value.

Strangers will appear even in a relatively closed environment. The presence of stranger features would lead to an estimated number of regular persons that is higher than the true value. We propose an algorithm to deal with this problem. Figure 5a illustrates *k*-means clustering result with 3 centers. The stranger feature is assigned to cluster ○. Evaluation sore of this result is 0.8. Figure 5b illustrates the result of clustering with 4 centers. The stranger feature forms an

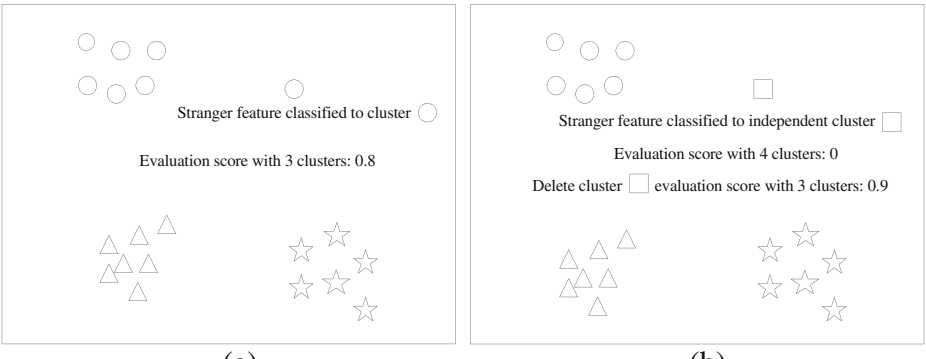

Fig. 5 Illustration when stranger feature forms an independent cluster

independent cluster □. This cluster has few members in it. We regard it as stranger feature and delete this cluster. Now we have 3 clusters. The evaluation score of this clustering (after deleting stranger feature) is 0.9. We choose the clustering result with the highest score as the optimal clustering. So for data in Fig. 5, the 3 clusters in Fig. 5b, after deleting the independent cluster □, are the optimal clustering results. The quality evaluation method is summarized in Algorithm 2. Steps 5~7 are designed to remove the impact of the stranger features. *Score* $[n{=}n_{actual}]$ is the clustering score with n_{actual} centers (without stranger features deleted).

The above evaluation method is based solely on the data in one *group (C, V)*. We need to consider all groups for decision making. Equation (15) can be replaced by:

$$n_{best} = \underset{n}{\operatorname{argmax}} \left(\sum_{C,V} Score(results(data_{CV}, n)) \right) \tag{20}$$

where $data_{CV}$ are the data in *group (C, V)*.

Algorithm 2: Clustering quality evaluation in one group

Input:

Data in the group, the interval of the number of regular persons $[n_1, n_2]$

Output:

An array of the clustering quality

1: Allocate an array *Score* $[n_1{:}n_2]$ for the clustering quality

2: Initialize the whole array *Score* with zeroes

3: **for** $(n{=}n_1; n{<=}n_2; n{++})$ **do**

4: Cluster the data in the group, the number of clusters is set to n

5: Delete clusters that contain few members and the number of clusters is changed from n to n_{actual}

6: Compute the clustering score S (with stranger features deleted) with cluster number n_{actual} according to Eqs. (16) and (19)

7: *Score*=max (*Score* $[n{=}n_{actual}]$, S)

8: **end for**

Establishing the correspondence between the New gallery and the Old gallery The gallery defined in Eq. (11) can be written in the following form:

$$\begin{aligned} gallery &= \{P_1, P_2, \cdots, P_{N_P}\} \\ P_i &= \left\{A_{P_i,C,V} | i \in [1, N_P], C \in [1, N_C], V \in [1, N_V]\right\} \end{aligned} \tag{21}$$

Assume that the old *gallery* is represented by Eq. (22) and the newly constructed *gallery* represented by Eq. (23). Note that the number of persons in the old *gallery* is N_P, which is not necessarily equal to its counterpart N_P' in the new *gallery*.

$$OldGallery = \{P_1, P_2, \cdots, P_{N_P}\} \tag{22}$$

$$NewGallery = \left\{ P_1{}', P_2{}', \cdots, P_{N_{P'}}{}' \right\} \tag{23}$$

As discussed above, P_i and P_i' do not necessarily represent the same person. *Correspondence* should be established among the members in the old *gallery* and those in the new *gallery* Eq. (24):

$$Correspondence = \left\{ (i,j) \middle| SamePerson\left(P_i, P_j{}'\right) = true \right\} \tag{24}$$

The function *SamePerson* is the process to estimate if the two objects are the same one, which is outlined in steps 4~9 in Algorithm 3.

In order to establish the correspondence between the old *gallery* and the new *gallery*, we need to calculate the similarity among their members. We construct a matrix $\mathbf{similarity}_{NpNp}$,whose elements satisfy:

$$similarity_{ij} = Similarty\left(P_i, P_j{}'\right) \tag{25}$$

$$Similarity\left(P_i, P_j{}'\right) = \frac{\sum_{C,V} Similarity\left(A_{P_i,C,V}, A'_{P'_j,C,V}\right)}{N_{combination}} \tag{26}$$

$A_{P_i,C,V}$ is the appearance feature vector of person P_i in camera C and viewpoint V in the old gallery. $A'_{P'_j,C,V}$ is the appearance feature vector of person P_j' in the same camera and viewpoint in the new gallery. $N_{combination}$ is the number of combinations of camera C and viewpoint V which observe the appearance model of both A_{P_i} and $A'_{P'_j}$. $N_{combination}$ is usually smaller than $N_C \times N_V$. Algorithm 3 outlines the correspondence approach. Steps 8 and 9 ensure that once a person in the *Oldgallery* corresponds to a person in the *Newgallery*, it cannot correspond to any other persons.

Once a target person O is being tracked, we obtain its camera identity C and viewpoint identity V. Re-identification is done by comparing the appearance of the target person A_O with the appearance model of all persons in camera C and viewpoint V in the gallery (see Eqs. (27) and (28)). Note that the topology information of the camera network is not used in re-identification.

$$ID_{P\,MaxSimlarity} = \underset{ID_P}{\operatorname{argmax}} \left(Similarity(A_O, gallery(ID_P, C, V)) \right) \tag{27}$$

$$ID_O = \begin{cases} ID_{P\,MaxSimlarity} & ifMaxSimilarity > \theta \\ stranger & otherwise \end{cases} \tag{28}$$

where ID_O is the identity of the target person O, and θ is the similarity threshold we choose when we can get the best performance on this dataset. *Similarity* is the similarity function defined in Eq. (6).

Algorithm 3: Establishing the correspondences between the person in the old gallery and that in the new gallery

Input:
Old gallery and new gallery
Output:
The new gallery aligned with the old one
1: Construct the similarity matrix **similarity**$_{NpNp}$ according to Eqs. (25), and (26)
2: Set a similarity threshold θ
3: **loop**
4: Find the maximum element $similarity_{max}$ in **similarity**$_{NpNp'}$
5: Record row number i_0 and column number j_0 of the maximum element
6: **if** $similarity_{max} > \theta$ **then**
7: The j_0-th person in the new gallery is the i_0 -th person in the old gallery
8: All elements in the i_0 -th row in **similarity**$_{NpNp}$ are set to zero
9: All elements in the j_0 -th column in **similarity**$_{NpNp}$ are set to zero
10: **else**
11: Break loop
12: **end if**
13:**end loop**

3.4 Time complexity analysis

For the static gallery learning in section 3.2, the main computational cost is *k*-means clustering in each group. The time complexity for *k*-means clustering is $O(K \cdot N \cdot I)$, where K is the number of clusters, N is the number of data, and I is the number of iterations required for convergence. Since *k*-means clustering is implemented in each *group (C, V)*, the time complexity can be estimated by:

$$O\left(\sum_{C,V} N_P \cdot N_{C,V} \cdot I\right) \tag{29}$$

where N_P is the number of regular persons, $N_{C,V}$ is the number of data in *group (C,V)*. Equation (29) can be further simplified to:

$$O\left(\sum_{C,V} N_P \cdot N_{C,V} \cdot I\right) = O\left(N_P \cdot I \cdot \sum_{C,V} N_{C,V}\right) = O(N_P \cdot I \cdot N_d) \tag{30}$$

where N_d is the total number of data involved in static gallery learning.

For online gallery learning, the gallery updates when a new track is recorded. For each updating, the time complexity lies in estimating the number of persons using different clusters chosen from $[n_{best}\text{-}n_{half}, n_{best}+n_{half}]$. The time complexity can be denoted as:

$$O\left(I \cdot N_d \cdot \sum_{n_{test}=n_{best}-n_{half}}^{n_{best}+n_{half}} n_{test}\right) \tag{31}$$

where n_{best} is the estimated number of regular persons before this updating. n_{half} is roughly half of the estimating width n_{width}. Equation (31) can be simplified to:

$$O\left(I{\cdot}N_d{\cdot}\sum_{n_{test}=n_{best}-n_{half}}^{n_{test}=n_{best}+n_{half}} n_{test}\right) = O\left(I{\cdot}N_d{\cdot}\left(2n_{half}+1\right){\cdot}n_{best}\right) = O(I{\cdot}N_d{\cdot}n_{width}{\cdot}n_{best}) \tag{32}$$

The computational cost for person re-identification lies in the comparison of the target person and the appearance models in the gallery. The time complexity is $O(N_p)$, where N_P is the number of regular persons.

4 Experiments and analysis

In this section, we present the experimental results of the proposed method in two different multi-camera scenarios. The scenarios different from each other both in terms of camera topologies and scene illumination conditions, and include both indoor and outdoor settings. In both scenarios we use the method proposed in [30, 36] to separate the foreground from the background. Foreground regions with large area are treated as moving objects. The bounding box of the corresponding foreground area is treated as the object's bounding box. In order to get the correct track that belongs to the same object, we associate different bounding boxes based on the appearance similarity of the objects and the physical positions of the bounding boxes. Bounding boxes belonging to the same object form the track of theObject. We also use some strategies to detect occlusions. When occlusions occur we did not extract appearance features.

We use the learned gallery to re-identify persons. The accuracy depends on two factors: 1) whether the target person is a regular person and 2) if the target is a regular person, confirm his/her identity in the gallery. The first factor is determined by the similarity threshold θ. The second factor is determined by the quality of the gallery. Re-identification accuracy and the wrong number of decisions made during evaluating the two factors are given in the experimental results.

The indoor dataset is used to assess the feasibility of our method. The outdoor dataset incorporates much noises compared with the indoor dataset, and we use it to test the robustness of the proposed method. In order to demonstrate the effectiveness of dynamic gallery learning, it is compared with the static gallery learning method [33] and two invariant feature based methods: SDALF [9] and FSCH [32].

4.1 Indoor experiments

The indoor dataset is composed of four video sequences captured by four non-overlapping cameras mounted inside a building. The entrance/exit of a camera and the hidden path that connects an entrance-exit pair are estimated using the method proposed in [6, 26]. It has to be pointed out that, an entrance is also an exit when an object leaves the FoV through it and vice versa. The estimated topology information of the indoor camera network is shown in Fig. 6. The ellipses denote the entrances/exits, and the straight lines denote the hidden paths that connect an entrances-exits pair. Each path has a time gap attached. The topology information can be learned and updated online. We load the static topology information and disable the updating of it in order to concentrate mainly on dynamic gallery learning.

There are four regular persons in the indoor camera network (see Fig. 7). A total of 438 tracks were obtained and 28 of them belong to the strangers. From Fig. 7, we can see that: 1)

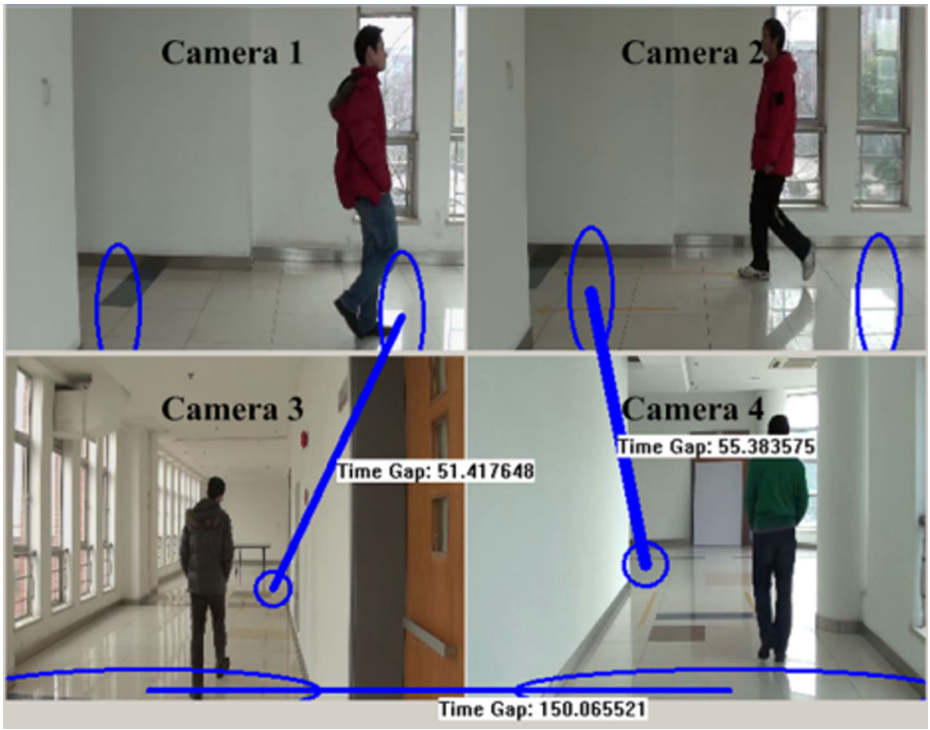

Fig. 6 Camera topology information of the indoor scenario. The ellipses denote the entrances/exits. The blue lines denote that there is a hidden path connecting the entrance-exit pair

person 1 and person 2 look very similar in appearance in camera 1 and 2, 2) the global illumination in camera 3 is much better than that in the other cameras and 3) people only present their left and right views in cameras 1 and 2, and front and back views in cameras 3 and 4. Viewpoints are decided by the directions of the paths each camera covers.

The first experiment is carried out to see whether our system can obtain the true number of regular persons. After initialization, the feature pool keeps collecting data until reaches its

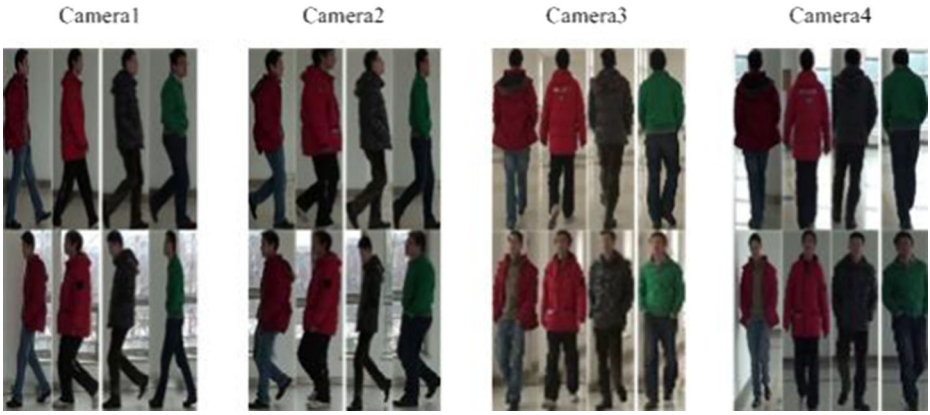

Fig. 7 Sample images of different objectss in each camera and viewpoint

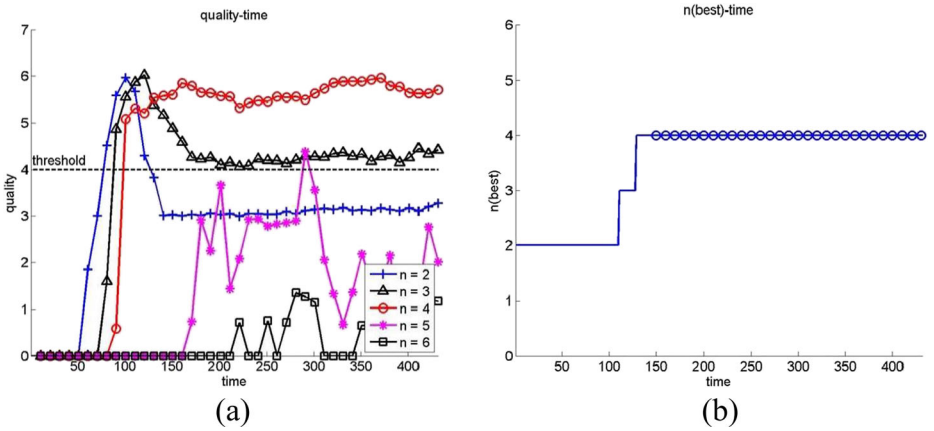

Fig. 8 Illustration of decision making on the number of regular persons in the indoor dataset. **a** Clustering qualities. **b** The best number of clusters versus time

maximum size. Let n be the tested number of regular persons (number of clusters). Figure 8a shows the curves of clustering quality (evaluation score) using different n versus time. The similarity threshold described in section 3.3 is also depicted. Note that the axis "time" is not the actual time, but rather the number of tracks. We use the axis "time" because our system is data triggered. As shown in Fig. 8a, all curves have quality of zero at the beginning. As time goes by, the curves corresponding to n=2, n=3 and n=4 get the highest quality (among all curves at the same time). Finally, the curve corresponding to n=4 gets a stable highest quality, which is our desired result because the true number of regular persons is 4. Let n_{best} be the best estimated number of regular persons. Figure 8b shows the curve of n_{best} versus time. We can see that it roughly follows the quality curve which gets the highest quality shown in Fig. 8a. The small circles in Fig. 8b indicate the time when the gallery is updated (sampled every 10 points to get a clear view). The locations of these circles indicate that the main phase of the gallery learning module runs only when n_{best} stabilizes for a pre-defined period of time and its clustering quality is higher than the quality threshold.

As for the correspondence between persons in the old gallery and those in the new gallery, no correspondence error occurs. We use the dynamic gallery for re-identification. The results are shown in Table 1, where column "Features" is the features used in each method. For dynamic gallery learning and static gallery learning, besides the FSCH feature stated before, we also test its combination with the height-width aspect ratio (AR) of the object. In this case

Table 1 Re-identification results of indoor experiment

Methods	Features	Best θ	Tracks evaluated	Correct decisions	Incorrect type 1	Incorrect type 2	Accuracy (%)
Dynamic gallery	FSCH	0.6	221	212	8	1	95.93
Dynamic gallery	FSCH+AR	0.75	221	217	4	0	98.19
Static gallery [33]	FSCH	0.35	221	210	10	1	95.02
Static gallery [33]	FSCH+AR	0.55	221	216	5	0	97.73
Invariant feature [32]	FSCH	0.25	221	172	20	29	77.83
Invariant feature [9]	SDALF	0.20	221	169	25	27	76.47

we use the "FSCH+AR" feature for both gallery learning and re-identification. We get the AR of each object using the bounding boxes of him/her in each camera and viewpoint. Column "Best θ" is the similarity threshold that produces the best re-identification result on this dataset (θ controls the decision of the designation of a stranger or a regular person (see Eq. (28)). Column "Tracks evaluated" is the number of tracks evaluated in each experiment. "Incorrect type 1" is the number of wrong decisions made during estimating whether the target is a regular person (factor 1), "Incorrect type 2" is the number of wrong decisions made during validating the identity of the target (factor 2). The number of incorrect type 2 reflects the quality of the learned gallery. "Accuracy" is the re-identification result corresponding to best θ.

In static gallery learning, all the 438 tracks are divided into a training set and a testing set. The training set is used to learn a static gallery, which is used to re-identify persons in the testing set. As a result, a total of 221 tracks are used for testing. To make fair comparisons, we use the same 221 tracks to test the performance of dynamic gallery learning, FCSH and SDALF. For our dynamic gallery learning method, 150 tracks from the training set are used to learn an initial gallery. For FSCH and SDALF, we manually create a gallery with regular person represented by his/her average FSCH or SDALF feature, regardless of the camera identity or viewpoint identity.

From Table 1 we can see that, when using the FSCH only, the accuracy using dynamic gallery and static gallery is 95.93 % and 95.02 % respectively. When combining FSCH with AR, the accuracies reach to 98.19 % and 97.73 %. Dynamic gallery learning is slightly better in both cases. As the illumination does not change significantly for the indoor dataset, the number of incorrect type 2 is the same for both methods. These results indicate that our dynamic gallery learning does not reduce the quality of the gallery. In addition, dynamic gallery learning does not need human intervention. The gallery can be learned and updated while re-identification is in progress. Dynamic gallery learning significantly outperforms FSCH and SDALF. The reason is that, FSCH and SDALF focus on features that are invariant and discriminative to describe a person, regardless of the differences in cameras and viewpoints. The features discriminative in one setting may not be appropriate for distinguishing objects in another setting. Unlike FSCH and SDALF, our method seeks to build a gallery to facilitate re-identification. The gallery contains the appearance models of an object in different cameras and viewpoints. Re-identification is performed by comparing the target and the appearance models in the corresponding camera and viewpoint. Even so, FSCH and SDALF are valuable since they can be used in any scenario. Meanwhile, our gallery learning approach should be used in a relatively closed environment where most objects appear frequently.

4.2 Outdoor experiments

The outdoor dataset consists of four videos taken from an outdoor environment (see Fig. 9). The dataset contains 931 tracks and 38 of them belong to the strangers. The number of regular persons is 10. Five regular persons captured in different cameras and viewpoints are shown in Fig. 10. Since most objects appear frequently in the camera network, the outdoor environment can also be viewed as a relatively closed environment.

This dataset is more challenging than the indoor dataset for three reasons. First, three regular persons leave the camera network for a long time and then re-enter. Second, the illumination changes significantly over time in each camera view, as can be seen in Fig. 9a and b. Finally, more regular persons exist in the outdoor dataset.

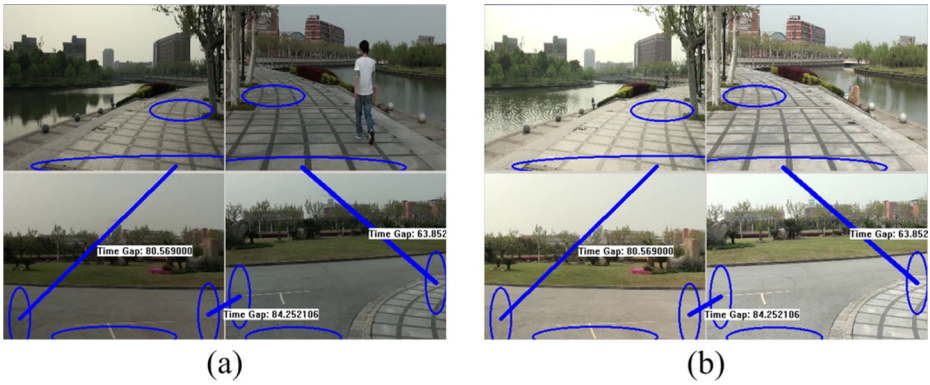

Fig. 9 Camera topology information of the outdoor scenario. **a** Dark scenes. **b** Bright scenes

Illumination variations have a great impact on our online gallery learning and updating. There are two categories of illumination variations in the outdoor dataset: illumination variations between cameras and illumination variations in each camera view over time. Persons' appearances in camera 1 are obviously darker than those in camera 2 because of the backlight problem (see Fig. 10), and the illumination in each camera changes significantly over time (see Fig. 9). Our dynamic gallery can deal with the first kind of variations since it has different appearance models in different cameras. It has to be pointed out that, to some extent, the overwriting mechanism of the feature pool can compensate for the light changes over time. However, if we do not adjust the brightness and rely solely on the overwriting mechanism, frequent illumination changes over time would significantly prevent the clustering progress from getting to a stable state. There would be no adaptation of the gallery if the clustering is unstable. To better deal with the second type of illumination variation, we use the brightness of a static background as the standard and adjust the brightness of the following images to the same value [23].

Clustering quality and the best estimated number of regular persons are shown in Fig. 11. The small circles in Fig. 11b denote the time that the gallery is updated. Both the curves and the marked points are sampled every 20 points to get a clear view.

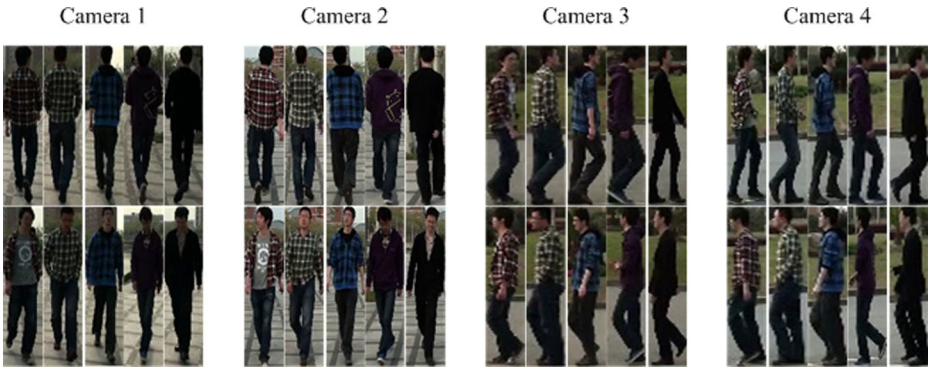

Fig. 10 Sample images of five regular persons in different cameras and viewpoints

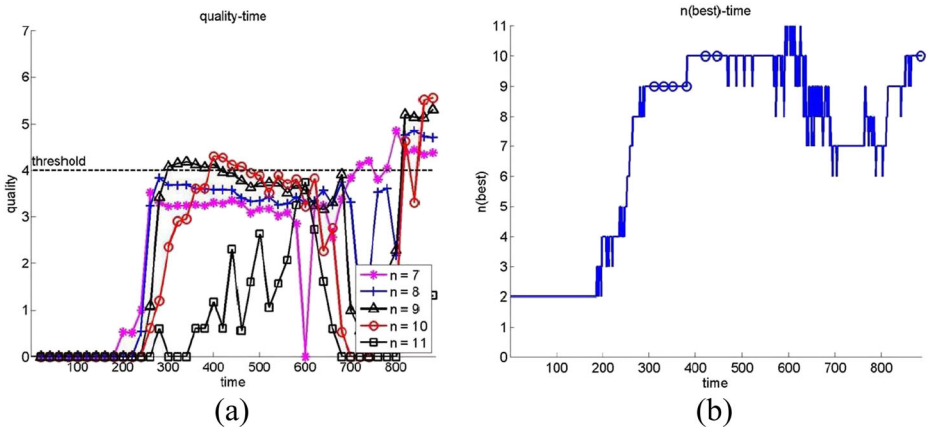

Fig. 11 Illustration of decision making on the number of regular persons in the outdoor dataset. **a** Clustering qualities. **b** The best number of clusters versus time

At the beginning, n_{best} is manually set to 2 (see Fig. 11b). The true number of regular persons is 10, which is not included in the test range $[n_1, n_2]$ ([9, 11]). From Fig. 11b we can see that n_{best} gradually moves to n_{true} (the true number of regular persons). There is a period that n_{best} stays to be 9. This is because the appearance of person 1 and person 2 are similar (see Fig. 10), and they are classified into the same cluster. The gallery obtained in this period is incorrect. n_{best} reaches to 10 with the introduction of more data and remains stable for a long time except for a few fluctuations. In this period, the gallery is correctly updated. After that, 3 regular persons leave the camera network and their features are gradually replaced by the remaining 7 objects' feature. Therefore, n_{best} falls to 7 after a period of fluctuations. Finally, the 3 objects re-enter to the camera network and n_{best} is updated to 10 again. Note that if we set a larger size limit of the feature pool, the system may "think" that no object has ever left. The setting of the size limit depends on the intention of the user.

The outdoor dataset contains 931 tracks. For static gallery learning and the two invariant feature based methods, the first 470 tracks are used as the training set to learn a gallery. The remaining 461 tracks are used as the testing set. Re-identification results using different methods in the outdoor dataset are given in Table 2. From this table we can see that, whether using feature "FSCH" or "FSCH+AR", dynamic gallery learning performs better than static gallery learning.

Table 2 Re-identification results in outdoor dataset

Methods	Features	Best θ	Tracks evaluated	Correct decisions	Incorrect type 1	Incorrect type 2	Accuracy (%)
Dynamic gallery	FSCH	0.4	461	427	23	11	92.63
Dynamic gallery	FSCH+AR	0.6	461	433	13	15	93.93
Static gallery [33]	FSCH	0.35	461	426	10	25	92.41
Static gallery [33]	FSCH+AR	0.6	461	430	8	23	93.27
Invariant feature [32]	FSCH	0.1	461	306	13	142	66.38
Invariant feature [9]	SDALF	0.35	461	319	35	107	69.19

There are three regular persons leave and re-enter in the camera network in the 461 testing tracks. As the static gallery of regular persons has already been learned using the training set, the change of regular persons have little impact on static gallery. The dynamic gallery is data triggered, it has to dynamically estimate the number of regular persons and update the gallery. As a result, the change of regular persons has more impact on dynamic gallery. So the number of incorrect type 1 is higher for dynamic gallery learning than static gallery learning. However, the number of incorrect type 2 is the lowest for dynamic gallery learning, indicting a higher quality gallery. These can be attributed to the fact that, dynamic gallery learning can adapt to the illumination variations in the outdoor environment. For static gallery learning, the appearance models remain unchanged over time. Illumination changes will cause mismatch between the target and the appearance models in the gallery. Our gallery learning framework, whether static or dynamic, outperforms the invariant feature based methods: FSCH and SDALF. These results bear out the benefit of introducing a multi-view gallery for person re-identification.

4.3 Runtime analysis

Table 3 illustrates the average time to update the gallery of our dynamic gallery learning method in both indoor and outdoor experiments. Column "actual ratio" is ratio between the actual time cost in the indoor experiment and the outdoor experiment. While column "theoretical ratio" is the ratio computed using Eqs. (30) and (32).

All experiments are implemented in Visual C++ platform with an Intel 2.67GHz CPU. Feature extraction and single camera tracking are not included in time measurement. The whole system works in real time. The codes could be optimized to save about half of the computational cost. Parallel computing using multiple CPU cores or GPU would further reduce the average updating time.

According to Eq. (32), the computational cost of dynamic gallery learning is $O(I \cdot N_d \cdot n_{width} \cdot n_{best})$, where I is the number of iterations of k-means clustering, N_d is the number of data, n_{width} is the estimating width, and n_{best} is the best estimated number of regular persons before this updating. In the indoor experiments, the number of data is 220. The best estimated number of regular persons remains to 4 for a long period (see Fig. 8b). In the outdoor experiments, the number of data is 440. The best estimated number of persons remains to 10 for a long period (see Fig. 11b). The number of iterations I is the same for both experiments. According to Eqs. (30) and (32), the theoretical ratio of updating times for both experiments is 1:5. The actual ratio is 1:3.14. The difference of the two ratios lies in that the best estimated number of regular persons is not a constant value. We use a constant value to calculate the theoretical ratio.

The time complexity of re-identification is $O(N_p)$, where N_P is the number of regular persons in the gallery. According to Table 3, the actual ratio and the theoretical ratio are approximately the same. This is because the re-identification process does not involve any unpredictable variables.

Table 3 Average Run Time of our Online Gallery Learning Framework

Methods	Indoor experiments	Outdoor experiments	Actual ratio	Theoretical ratio
Online Gallery learning	81.75 ms	257 ms	1 : 3.14	1 : 5
Re-identification	0.0061 ms	0.0152 ms	1 : 2.49	1 : 2.5

5 Conclusions

In this paper, we proposed a multi-view gallery learning method for person re-identification, regardless of the specific features adopted. The gallery contains the appearance models of the regular persons in a relatively closed environment. Since the appearance models are accumulated from different cameras and viewpoints for each object, they are robust to illumination changes, pose variations and camera settings. With the learned gallery, re-identification becomes the problem of finding the target from the gallery in the corresponding camera and viewpoint. In addition, our gallery learning is dynamic and online.

Compared with static gallery, dynamic gallery is more adaptive to illumination variations over time, and it is more convenient in application as it does not need a training phase and human intervention. The computational cost of each updating of the gallery is inexpensive and has great potential for improvement, which facilitates real time applications. Our multi-view gallery learning method outperforms benchmark appearance based methods in both indoor and outdoor datasets. The main drawback of our method is that, it can only be applied to a relatively closed environment. The appearance based methods do not have such a limitation.

Currently, the appearance models in the gallery preserve the visual cues the regular person. As future works, the multi-view gallery method should learn a gallery that combines multiple types of features, like gait and topology information. With the combined features, the gallery should be able to deal with objects wearing similar clothes and objects changing their clothes. Besides, the viewpoint of an object is evaluated by his moving direction. When the object keeps still, the approach cannot get his viewpoint. This problem can be addressed using other techniques in computer vision, like viewpoint estimation.

Acknowledgments This research has been partially supported by the grants of China 973 project 2011CB302203, NSFC 61375019 and NSFC 61273285.

References

1. Bak S, Corvee E, Bremond F and Thonnat M (2010) "Person Re-identification Using Spatial Covariance Regions of Human Body Parts," in Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 435–440
2. Bak S, Corvee E, Bremond F, and Thonnat M (2010) "Person re-identification using Haar-based and DCD-based signature," in Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp.1-8
3. Bak S, Corvee E, Bremond F, and Thonnat M (2011) "Multiple-shot human re-identification by Mean Riemannian Covariance Grid," in Proc. of IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 179–184
4. Bak S, Corvee E, Bremond F, Thonnat M (2012) Boosted human re-identification using riemannian manifolds. Image Vis Comput 30(6):443–452
5. Bazzani L, Cristani M, Perina A, Murino V (2012) Multiple-shot person re-identification by chromatic and epitomic analyses. Pattern Recogn Lett 33(7):898–903
6. Chen K, Lai C, Hung Y and Chen C (2008) "An Adaptive Learning Method for Target Tracking across Multiple Cameras," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 1–8
7. Cheng D, Cristani M, Stoppa M, Bazzani L, and Murino V, (2011) "Custom pictorial structures for re-identification," in Proc. of British Machine Vision Conference (BMVC)
8. Dikmen M, Akbas E, Huang T, and Ahuja N (2010) "Pedestrian recognition with a learned metric," in Proc. of Asian Conference on Computer Vision (ACCV), pp. 501–512

9. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2013) Symmetry-driven accumulation of local features for human characterization and re-identification. Comput Vis Image Underst 117(2):130–144
10. Gandhi T, Trivedi M (2007) Person tracking and reidentification: introducing panoramic appearance Map (PAM) for feature representation. Mach Vis Appl (MVA) 18(3):207–220
11. Gheissari N, Sebastian T, and Hartley R (2006) "Person reidentification using spatiotemporal appearance," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1528–1535
12. Gray D and Tao T (2008) "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in Proc. of European Conference on Computer Vision (ECCV), pp. 262–275
13. Hamdoun O, Moutarde F, Stanciulescu B, and Steux B (2008) "Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences," in International Conference on Distributed Smart Cameras (ICDSC), pp. 1–6
14. Hirzer M and Beleznai C and Roth P and Bischof H (2011) "Person re-identification by descriptive and discriminative classification," Image Analysis, pp. 91–102
15. Javed O, Shafique K, and Shah M (2005) "Appearance modeling for tracking in multiple non-overlapping cameras," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 26–33
16. Jeong K, Jaynes C (2008) Object matching in disjoint cameras using a color transfer approach. Mach Vis Appl (MVA) 19(5–6):443–455
17. Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. John Wiley & Sons, Hoboken
18. Kostinger M, Hirzer M, Wohlhart P, Roth P, and Bischof H (2012) "Large scale metric learning from equivalence constraints," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2288–2295
19. Kviatkovsky I, Amit A, Rivlin E (2013) Color invariants for person reidentification. IEEE Trans Pattern Anal Mach Intell (PAMI) 35(7):1622–1634
20. Li Wand Wang X (2013) "Locally Aligned Feature Transforms across Views," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3594–3601
21. Li Z, Chang S, Liang F, Huang T, Cao L and Smith J (2013) "Learning Locally-Adaptive Decision Functions for Person Verification," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3610–3617
22. Li W, Zhao R, and Wang X (2012) "Human reidentification with transferred metric learning," in Proc. of Asian Conference on Computer Vision (ACCV), pp. 31–44
23. Loy C, Xiang T, Gong S (2010) Time-delayed correlation analysis for multi-camera activity understanding. Int J Comput Vis 90(1):106–129
24. Ma B, Su Y, and Jurie F (2012) "Local descriptors encoded by fisher vectors for person re-identification," in Proc. of European Conference on Computer Vision Workshops and Demonstrations, pp. 413–422
25. Ma B, Su Y, and Jurie F, (2012) "Bicov: a novel image representation for person re-identification and face verification," in Proc. of British Machine Vision Conference (BMVC)
26. Makris D and Ellis T (2003) "Automatic Learning of an Activity-based Semantic Scene Model," in Proc. of IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 183–188
27. Mignon A, and Jurie F (2012) "PCCA: A new approach for distance learning from sparse pairwise constraints," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2666–2672
28. Prosser B, Zheng W, Gong S, and Xiang T (2010) "Person re-identification by support vector ranking," in Proc. of British Machine Vision Conference (BMVC), pp. 1–11
29. Schwartz W and Davis L (2009) "Learning discriminative appearance-based models using partial least squares," in Proc. of Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), pp. 322–329
30. Stauffer C and Grimson W (1999) "Adaptive Background Mixture Models for Real-time Tracking," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
31. Wang X, Doretto G, Sebastian T, Rittscher J, and Tu P (2007) "Shape and appearance context modeling," in Proc. of IEEE International Conference on Computer Vision (ICCV), pp. 1–8
32. Xiang Z, Chen Q, and Liu Y. (2012)"Person re-identification by fuzzy space color histogram," Multimedia Tools and Applications, pp. 1–17, 2012
33. Xiang Z, Chen Q, and Liu Y (2013) "Feature correspondence in a non-overlapping camera network," Multimedia Tools and Applications, pp. 1–17

34. Zhao R, Ouyang W, and Wang X, (2013) "Unsupervised Salience Learning for Person Re-identification," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
35. Zheng W, Gong S, and Xiang T (2011) "Person Re-identification by Probabilistic Relative Distance Comparison," in Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 649–656
36. Zivkovic Z (2004) "Improved Adaptive Gaussian Mixture Model for Background Subtraction," in Proc. of IEEE International Conference on Pattern Recognition (ICPR), vol. 2, pp. 28–31

Yanna Zhao received the M.S. degree in signal and information process from Shandong Normal University, Jinan, China, in 2010. She is currently a PhD candidate at School of Information Science and Engineering, Shandong University, Jinan, China. She is now doing research in Institute of Image Processing and Pattern Recognition at Shanghai Jiao Tong University. Her research interests include computer vision, pattern recognition and image processing.

Xu Zhao received the Ph.D. degree in pattern recognition and intelligence system from Shanghai Jiao Tong University in 2011. He is currently an Associate Professor in the department of Automation at Shanghai Jiao Tong University. He was a visiting scholar at the Beckman Institute for Advanced Science and Technology at University of Illinois at Urbana-Champaign from 2007 to 2008. He had been the postdoctoral research fellow in the Northeastern University from 2012 to 2013. His research interests include visual analysis of human motion, machine learning and image/video processing.

Zongjie Xiang received his MS degree at 2009 in Shanghai University, China. He got his PhD degree at 2014 at the institute of Image Processing and Pattern Recognition in Shanghai Jiao Tong University, Shanghai, China. His research interests include pattern recognition and computer vision.

Yuncai Liu received his Ph.D. in the Department of Electrical and Computer Science Engineering from the University of Illinois at Urbana-Champaign in 1990 and worked as an associate researcher at the Beckman Institute of Science and Technology from 1990 to 1991. Since 1991, he was a system consultant and then chief consultant of research at Sumitomo Electric Industries, Ltd., Japan. In October 2000, he joined Shanghai Jiao Tong University as a distinguished professor. His research interests are in image processing and computer vision, especially in motion estimation, feature detection and matching, and image registration. He also has made great progress in the research of intelligent transportation systems.