SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2018, Vol. 61 092108:1–092108:14
https://doi.org/10.1007/s11432-016-9086-8

Histogram of the node strength and histogram of the edge weight: two new features for RGB-D person re-identification

Zeynab IMANI & Hadi SOLTANIZADEH*

Electrical and Computer Engineering Faculty, Semnan Univercity, Semnan 46198*, Iran*

Received 17 January 2017/Accepted 16 March 2017/Published online 21 May 2018

Abstract Person re-identification is a classical task for any multi-camera surveillance system. Most of the existing researches on re-identification are based on features extracted from RGB images. However, there are many deficiencies in RGB image processing, some of which are requiring a lot of illumination and high computation. In this paper, novel features are proposed for RGB-D person re-identification. First, the complex network approach in texture recognition is modified and its threshold function is changed for using in depth images extracted by RGB-D sensors. Then, two novel measurements named the histogram of the edge weight (HEW) and the histogram of the node strength (HNS) are introduced on complex networks. Our features fit both single-shot and multi-shot person re-identification. In the single-shot case, the HNS is extracted from only one frame while for the multi-shot case it is extracted from both one frame and multi-frames. These proposed measurements are called histogram of the spatial node strength (HSNS) and histogram of the temporal node strength (HTNS) respectively. Subsequently, these measurements are combined with skeleton features using score-level fusion. The method is evaluated using two benchmark databases and the results show that ours outperforms some state-of-the-art methods.

Keywords complex network, edge weight, graph, node strength, person re-identification, skeleton features

Citation Imani Z, Soltanizadeh H. Histogram of the node strength and histogram of the edge weight: two new features for RGB-D person re-identification. Sci China Inf Sci, 2018, 61(9): 092108, https://doi.org/10.1007/s11432-016-9086-8

1 Introduction

Re-identification is the task of recognizing a person entering a camera's field of view and have been previously seen by a different camera, or by the same camera at a different time instance. This task is very important due to its extensive application in video surveillance, behavior analysis, object tracking, and target search in a collection of video sequences. The person re-identification is fundamentally challenging because of the large visual appearance changes caused by variations in view angle, illumination intensity, poses, background clutter, image resolutions and occlusion. These make the inter-personal variations more significant than intra-personal variations.

In most researches, person re-identification is regarded as a retrieval problem. Given an image/video of an unknown person as probe, first, representative features are extracted to describe the person and then similarities between the probe person and all samples in a gallery set are computed. Generally, a ranked list of all people in the gallery is created so that the higher ranked person is more likely corresponds

* Corresponding author (email: h_soltanizadeh@semnan.ac.ir)

to one in the probe set. Usually, person re-identification is done using the clothing appearance features extracted from RGB images. Performing re-identification based on only these features is difficult because different subjects often wear clothes with similar colors. To avoid the mentioned problems in the RGB image processing, we perform re-identification using depth and skeleton images extracted by RGB-D sensors.

This paper presents a novel method for short-term person re-identification. By short-term we mean recognizing people within short time frames; supposing that a person is wearing the same clothing in both the gallery and probe sets. It should be noted that in our tests, video frames are taken by using low cost sensor Kinect camera. First, depth images are preprocessed to remove noise. If any noise existed, we reduce noise by using some noise reduction methods. In the next step, the depth images are modeled as complex networks. Then, two novel measurements are introduced on the complex networks; histogram of the edge weight (HEW) and histogram of the node strength (HNS). The HNS is extracted as spatial from only one frame (single-shot) and temporal from multi-frames (single-shot). Since the human body is very flexible and deformable, and usually presents a number of different texture and shape patterns on the different body parts. The complex networks because of their flexibility and generality to represent any given structure (such as lists, trees, networks and images) can be acquire a reliable re-identification. In addition, due to the graph approach, they are more robust against scale and rotation changes and can re-identify persons even using the incomplete images, for example, images which lose the information of the legs due to walking or the information of the body different parts due to the noise. On the other hand, in our task, complex networks are employed to extract both the spatial and temporal features. The former obtains the appearance and structural features and the latter obtains appearance, structural, and motion features (such as features related to gait). Finally, these features are combined with skeleton features using the score-level fusion. The skeleton features are able to actually to obtain similar descriptors in different observations of the same target using the structural features. Thus, the proposed method obtains three types of appearance, motion and structural features and can be achieve a reliable re-identification which is validated using our experiments. This paper is organized as follows: In Section 2, related work is reviewed. In Section 3, the proposed features are discussed. Experiments and results are given in Section 4 and the conclusion is presented in Section 5.

2 Related work

Conventional methods such as face [1], gait [2] or silhouette [3] recognition have been widely used in human identity recognition. However, due to the low resolution of images/videos, it is difficult to apply these methods for person re-identification.

The most of existing methods for person re-identification, have focused on the development of sophisticated features to explain the visual appearance of targets [4]. In [5] a person re-identification algorithm was introduced to fully exploit region-based feature salience. First, the image of each person was divided into the upper part and the lower part. Subsequently, a part-based feature extraction algorithm was proposed to adopt different features for different parts. In [6], discriminative appearance-based models were proposed using Partial Least Squares (PLS) over texture, gradients and color features. In [7], Symmetry-Driven Accumulation of Local Features (SDALF) descriptor was proposed by exploiting axis symmetry and asymmetry and presenting each body part using a weighted color histogram, Maximally Stable Color Regions (MSCR) and Recurrent High-Structured Patches (RHSP). In [8], the combination of Biologically Inspired Features (BIF) and covariance descriptors were introduced for person re-identification and face verification. To extract the BIF, the magnitude images of the Gabor filters were extracted from different spatial scales (i.e., BiCov). Then a single band was formed by grouping the neighboring scale responses. In [9], a Custom Pictorial Structure (CPS) model including the head, chest, thighs and legs part descriptors was proposed for person re-identification. This model were fitted using color histograms and MSCR features. In [10] re-identification was performed using dense color histogram and dense SIFT features. Then, adjacency-constrained patch matching was used to build dense correspondence between image

pairs through an unsupervised saliency learning method. In [11] saliency matching was proposed based on patch matching in a unified structural RankSVM learning framework for person re-identification. In [12] a spatial pyramid-based statistical feature extraction framework as a unified pipeline was proposed for extracting and combining multiple statistical features for person Re-identification. In this task, five types of spatial pyramid-based statistical features, including spatial pyramid-based color histogram (spHist), spatial pyramid-based histogram of oriented gradient (spHOG) spatial pyramid-based local binary pattern (spLBP), spatial pyramid-based color names (spCNs), spatial pyramid based covariance feature (spCov), and combine them via multiple kernel local Fisher discriminant analysis (mkLFDA) were extracted from images. In [13] a novel local descriptor named quaternionic local ranking binary pattern (QLRBP) was proposed for person re-identification. The QLRBP is able to handle all color channels directly in the quaternionic domain and include their relations simultaneously.

Different from the above-mentioned approaches, the metric learning methods have been recently proposed in person re-identification systems very much. These methods are specifically used to select the most relevant features with supervised [14] or unsupervised [15] learning algorithms in order to person re-identification. In [14], person re-identification was performed by a supervised method named RankSVM using pairs of similar and dissimilar images. In [16], the Probabilistic Relative Distance Comparison (PRDC) approach was introduced using learning a metric in which the probability of an incorrect match having a small distance is less than that of a correct one. In [17], a semi-supervised, manifold ranking method was used to single-shot person re-identification. This work was limited as only configurations with a single gallery image per person and a single test image was only used. In [18], fusion the different features was proposed using a supervised strategy named multi-feature learning (MFL) that at least required a single image per person as training data. In [19], KISSME learning method was introduced to learn a distance metric from equivalence constraints of a statistical inference perspective. In [20] dual-regularized KISS (DR-KISS) metric learning was introduced to improve KISSME metric by reducing overestimation of large eigenvalues of the two estimated covariance matrices. In [21], Pairwise Constrained Component Analysis (PCCA) was proposed to learn a projection from high-dimensional input space into a low-dimensional space in which the distance between pairs of data points was adopting the desired constraints. In [22], a Set-Label model was proposed using deep belief network and neighborhood component analysis to improve the person re-identification performance. For a more survey on person re-identification, refer to [4,23]. In [24] a unified deep ranking framework was proposed for person re-identification using direct prediction the similarity of a pair of pedestrian images via joint representation learning.

Before presenting the Kinect sensor, there have been systems and algorithms for person re-identification which used RGB cameras. The availability of affordable RGB-D sensors, Microsoft Kinect, made it possible to use it in many applications such as person re-identification [25–28]. In addition to the low cost of these sensors, the RGB-D sensors are almost insensitive to shadows and illumination changes, provide additional 3D shape information, add new types of features to the feature space, provide robust background subtraction, which simplifies multiple people tracking and allows calibrated virtual views of the person to be created.

In [25], the Kinect sensor was used to extract 3D soft biometric cues such as geometric features that are invariant to appearance variations. Their method was considered on long term person re-identification in which a person's appearance may change over time. In [26], a multi-camera system was proposed for person re-identification based on the relative positions of joints extracted from the skeleton provided by the Kinect sensor. In this task, the fast re-identification method was performed using dissimilarity representation descriptors. In [27], two types of features were proposed for person re-identification that the first type contains 13 anthropometric measures extracted from the body joints and the second type contains a point cloud model of the person body. In [28], the appearance-based re-identification accuracy was increased by fusing the clothing appearance features with anthropometric measures using a dissimilarity-based framework. In [29], a texture-based signature named Skeleton-based Person Signature (SPS) including local descriptors around the joint positions of the skeleton of the Kinect sensor was proposed for more robust person re-identification in the presence of strong illumination changes. In [30], the

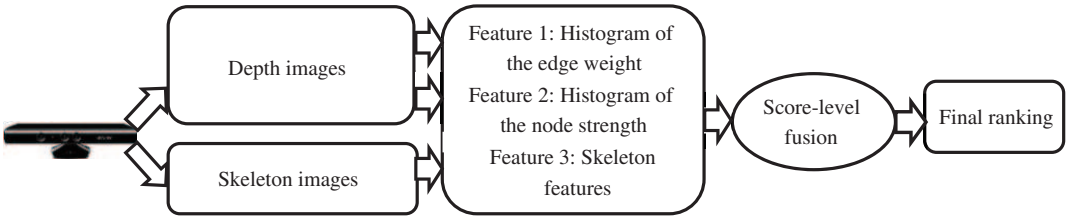

Figure 1 The proposed algorithm flowchart.

bodyprints were extracted from height maps in which the pixel values represent the height with respect to the ground. The key idea of bodyprints is that each of their elements summarizes the color appearance at a different height. Then, the latent features were generated using probabilistic latent variable models applied over the appearance descriptors of bodyprints in order to person re-identification. In [31] after converting depth images to point clouds, covariance descriptors were computed using 6-dimensional feature vectors including x-, y- and z-coordinates and surface normal vectors in 3 direction of x, y and z. Then, Eigen-depth features were extracted by employing the rotation transformation matrix on covariance descriptors in order to stability against rotation in different cameras and combined with skeleton features for RGB-D person re-identification.

3 Feature extraction

To extract features, first, depth images are preprocessed to remove noise. The depth images are divided into 25×25 blocks, and each block is tested for existence of noise. The depth values identified as noise are corrected using linear interpolation, that is assigning the average value of their 3×3 neighborhoods. In the next step, the depth images are converted to complex networks however with a different thresholding approach into previous tasks on complex networks. The complex networks are based on graph theory to convert the images to graphs and their evaluations are performed by measurements used on the graphs. Then, two novel measurements are introduced on the complex networks; the histogram of the edge weight (HEW) and the histogram of the node strength (HNS). For one image/frame (single-shot case), the histogram of the spatial node strength (HSNS) is extracted to fit the single-shot person re-identification case. In multi-images/frames (multi-shot case), both the histogram of the spatial node strength (HSNS) and the histogram of the temporal node strength (HTNS) are extracted to fit multi-shot person re-identification case. Finally, these features are combined with skeleton features using the score-level fusion. The flowchart of the proposed algorithm have been shown in Figure 1.

3.1 Complex networks

Complex network can be defined as the intersection between graph theory and statistical mechanics, which confer a truly multidisciplinary nature to this area [32]. Due to its great flexibility and generality, it has attracted a lot of attention in many areas such as biology [33], sociology [34], computer vision [32, 35, 36] and physics [37].

The majority of the existing research on complex networks includes two main steps: (a) representation of the problem as a complex network by the analysis of its topological features and (b) recognizing the different categories of network structures. The complex networks are represented by graphs. An undirected weighted graph $G = \langle V, E \rangle$ is defined by a set of nodes $V = \{v_i\}$ and a set of edges $E = \{e_{i,j}\}$, where $e_{i,j}$ represents the connection weight between nodes v_i and v_j.

Suppose that $I = (X, Y)$ is an image where $X = \{x_i\}$ is x-coordinate of pixels and $Y = \{y_i\}$ is y-coordinate of pixels. For modeling image, $I = (X, Y)$, as regular complex network, $G = \langle V, E \rangle$, first, each pixel is considered as a node of the graph. Then an edge is created between the nodes v_i and v_j if the Euclidean distance between their related pixels is equal to or lower than a given radius r. Subsequently, for each edge $e_{i,j} \in E$, a weight is defined based on difference of the gray level of pixels. In this task, the

depth value of pixels instead the gray level of pixels is used for calculating weight as following:

$$e_{i,j} = \begin{cases} |u_i - u_j|, & \text{if } d_{i,j} \leqslant r, \\ \text{NaN}, & \text{else}, \end{cases} \tag{1}$$

where u_i is the depth value of pixel i and $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ is the Euclidean distance between pixels of i and j.

Assume that $S = (X, Y, T)$ is a video including multi-frame where $X = \{x_i\}$ is x-coordinate of pixels, $Y = \{y_i\}$ is y-coordinate of pixels and $T = \{t_i\}$ is the number of the frames in the video. For modeling video, $S = (X, Y, T)$, as regular complex network, $G = \langle V, E \rangle$, first, each pixel is considered as the graph node. Then, an edge is created between the nodes of v_i and v_j if the Euclidean distance between their related pixels is equal to or lower than a given radius r. Subsequently, for each edge $e_{i,j} \in E$, a weight is defined as following:

$$e_{i,j} = \begin{cases} |u_i - u_j|, & \text{if } D_{i,j} \leqslant r, \\ \text{NaN}, & \text{else}, \end{cases} \tag{2}$$

where $D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (t_i - t_j)^2}$ is the Euclidean distance between pixels i and j.

For determining more image/video properties, a function is applied over the edges of regular complex network. This function removes edges according to their weights [38, 39]. The function is defined as remaining pixels with the edge weight equal to or lower from a given threshold of δ. This approach is not fully adequate to extract suitable features from depth images of persons because the depth images are built based on distance from sensor and whatever the distance between the depth values of two pixels increases, those pixels can be the distinguished points such as corners or peaks in image/video to detect more image/video properties. Therefore, an improved method to extract suitable features from complex networks is developed for depth images. The key idea of the method is to automatically select the edges with weight equal to or above a given threshold δ. In fact, we propose the inverse thresholding way into previous applications of complex networks in order to introduce the threshold function as following equation and opt edges with weight equal to or above δ.

$$e_{i,j} = \begin{cases} e_{i,j}, & \text{if } \ e_{i,j} \geqslant \delta, \\ \text{NaN}, & \text{else}. \end{cases} \tag{3}$$

Applying a set of thresholds δ to the original network can also be interpreted as the acquisition of various samplings of a complex network. This approach grants us a signature which removes temporary characteristics of the network and, therefore, a richer set of features can be extracted for describing the network behavior.

For each threshold δ, the regular complex network is transformed into a δ-scaled network. The δ-scaled network represents different properties compared with the regular complex network and determines the structure and topology related to its scale. In the low values of δ, whatever value r increases, more edge connections are obtained on the complex networks. Thus values of δ should increase by increasing radius r for obtaining more information. In the other words, values of δ and r should be traded off for achieving to complex networks with more effective information. Figure 2 shows complex networks with different values of δ and r for a subject randomly selected from KinectREID database.

The proposed measurements on complex networks. Each complex network represents specific topological features which specify its connectivity and highly affection in various applications adopted by complex network. The study of a complex network relies on the use of measurements capable of expressing the most relevant topological features. So far, many measurements have been introduced on complex networks such as degree, mean, contrast, energy, entropy and so on [40]. We introduce two new measurements named the histogram of the edge weight (HEW) and the histogram of the node strength (HNS). For the complex network model of image/video, the node strength (spatial node strength in the image case and temporal node strength in the video case) is the statistics to determine the amplitude

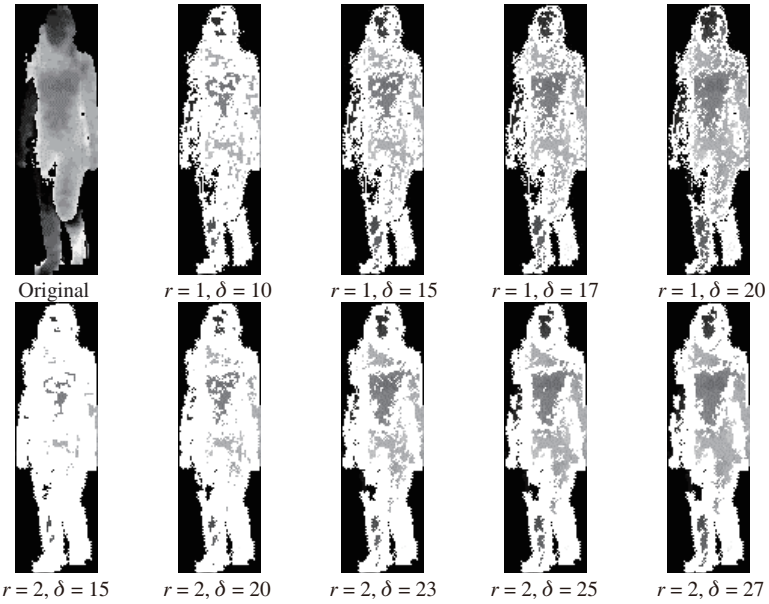

Figure 2 The depth images modelled as complex networks with different values of δ and r.

of changes in the depth values of pixels. It allows to study the structural features of the persons by using the complexity of the node connections in the network. The temporal node strength in addition to extracting the structural features, also obtains motion information especially about incomplete images for example about the legs part in which their information lose when persons gait. On the other hand, the edge weight reflects the depth relation among nodes and their neighbors which are basic elements for extracting the depth features. In addition, the histogram is obviously a simple and concise source of statistical information of the network topology. Based on above-mentioned advantages for node strength, edge weight and histogram, the HEW, HSNS and HTNS are much appropriate features for RGB-D person re-identification.

The spatial node strength, q_i^{spatial}, at node v_i, is defined as sum weights in each node in a given graph (in single-shot case) as following:

$$q_i^{\text{spatial}} = \sum_{v_j \in V} \begin{cases} e_{i,j}, & \text{if } e_{i,j} \in E, \\ 0, & \text{else}, \end{cases} \quad i = 1, \ldots, N, \tag{4}$$

where N is the number of the graph nodes.

Then, the HSNS, $\mathcal{H}_\nu^{\text{spatial}}$, is obtained as

$$\mathcal{H}_{\nu=q_i^{\text{spatial}}}^{\text{spatial}} = \sum_{v_j \in V, j \neq i} f\big(q_j^{\text{spatial}}, q_i^{\text{spatial}}\big), \quad \nu = 1, \ldots, M, \tag{5}$$

where M is the maximum value of q^{spatial} and $f(a, b)$ is calculated as following:

$$f(a,b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{else.} \end{cases} \tag{6}$$

However, the temporal node strength, q_i^{temporal}, at node v_i is sum weights in which node v_i in frame $t = \gamma_1$ connects to node v_j in frame $t = \gamma_k$ and is defined as

$$q_i^{\text{temporal}} = \sum_{t=\gamma_2}^{\gamma_T} \sum_{v_j \in V} \begin{cases} e_{i,j}, & \text{if } v_i \in \gamma_1,\ v_j \in \gamma_k,\ k = 2, \ldots, T,\ k \neq 1, \\ 0, & \text{else}, \end{cases} \quad i = 1, \ldots, N, \tag{7}$$

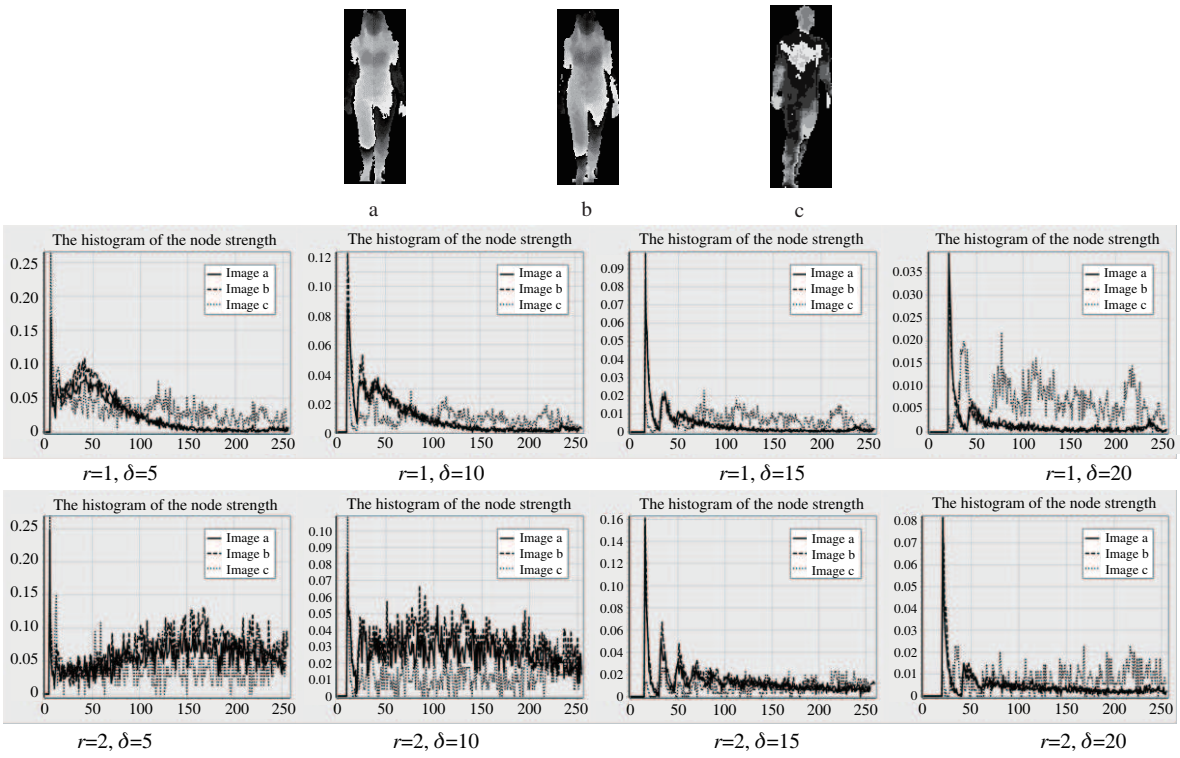

Figure 3 The HSNS features for three subjects (image a and image b belong to the same subject and image c belong to another subject).

where N is the number of the graph nodes and T is number of frames. Then, the HTNS, $\mathcal{H}_{\zeta}^{\text{temporal}}$, is obtained in a way similar to the HSNS as following:

$$\mathcal{H}^{\text{temporal}}_{\zeta=q_i^{\text{temporal}}} = \sum_{v_j \in V, j \neq i} f(q_i^{\text{temporal}}, q_j^{\text{temporal}}), \quad \zeta = 1, \ldots, M, \tag{8}$$

where M is the maximum value of q^{temporal}.

We also define the HEW, $\mathcal{H}_{\xi}^{\text{weight}}$, which connects node v_i to node v_j in a graph as following:

$$\mathcal{H}^{\text{weight}}_{\xi=e_{i,j}} = \sum_{e_{k,l} \in E} f(e_{k,l}, e_{i,j}) \forall \begin{Bmatrix} k \neq i,\ l \neq j \\ k = i,\ l \neq j \\ k \neq i,\ l = j \end{Bmatrix}, \quad \xi = 1, \ldots, S, \tag{9}$$

where S is the maximum value of weights in whole graph.

After extracting the histograms, the whole of them are normalized into interval $[0, 1]$ for invariance against the scale and rotation.

In the single-shot person re-identification case (i.e., in which specifically one sample image of each person is obtained in the gallery set and at least one instance of each person is obtained in the probe set), we employ features of the HEW and HSNS. In multi-shot case (i.e., in which a group of M samples from each subject is given in the gallery set, and a group of M samples from each subject is given to be re-identified in the probe set), we employ three types of the HEW, HSNS and HTNS features for re-identification. Figures 3 and 4 show the HSNS and HEW features for three subject respectively that image a and image b belong to the same subject and image c belong to another subject. It is seen from Figures 3 and 4 that the histogram features of images of the same subject have more similarity than those of the other subject.

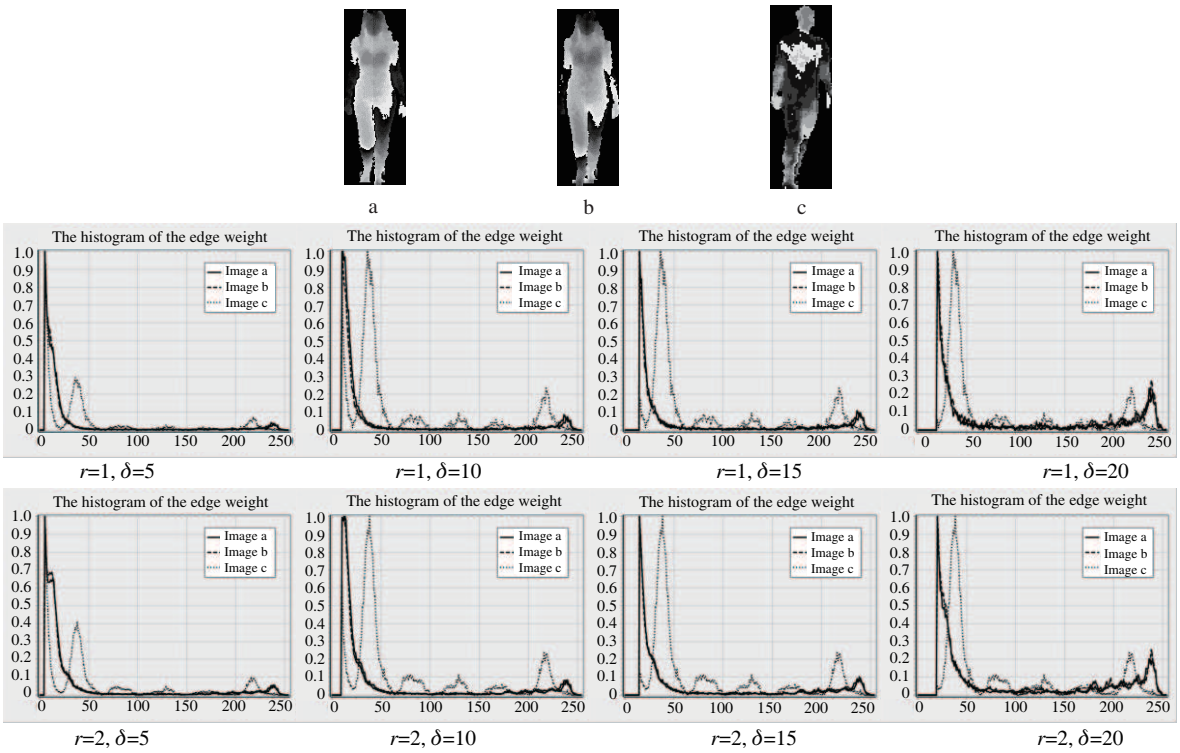

Figure 4 The HEW features for three subjects (image a and image b belong to the same subject and image c belong to another subject).

3.2 Skeleton features

The skeleton features are the characterization of subjects through the measurement of physical body features, e.g., height, arm length, and eye-to-eye distance [25,27,28]. The use of the skeleton features has been already proposed for person re-identification [25,28]. We use the skeleton features employed in [28] that contain distance between floor and head, the ratio between torso and legs (which is distance between neck and shoulder, division on the product of distance between floor and head; and distance between floor and hip), height (which is calculated as distance between the highest body silhouette point and the floor plane), distance between neck and shoulder, distance between floor and neck, distance between torso center and shoulder, distance between torso center and hip, arm length (which is sum of the distances between shoulder and elbow, and between the elbow and wrist), leg length (which is defined as sum of the distances between hip and knee, and between knee and ankle).

Then all the distances are normalized to a zero mean and unitary variance as in [28]. Subsequently the similarity between the skeleton features of two subjects X and Y is calculated using weighted Euclidean distance as in [25,28] where the weights present importance of features. The weights achieved in [28] are also used in this task.

4 Experiments and results

4.1 The evaluation method and the experimental setting

In re-identification, two sets of images/videos are available: the gallery set G and the probe set P. The re-identification consists matching each signature of the probe set P, I_P, to the corresponding signature of the gallery set G, I_G. The re-identification can be seen as a maximum log-likelihood estimation problem [41]. More in details, given a probe P the matching is performed as

$$G^* = \underset{G}{\operatorname{argmax}}(\log P(I_G|I_P)) \underset{G}{\operatorname{argmin}}(d(I_G, I_P)), \tag{10}$$

where $d(I_G, I_P)$ measures the distance between two features. In this task, the matching distance d is defined using the score-level fusion (fusing the matching scores calculated separately for each feature) that in the multi-shot person re-identification case is calculated as

$$\begin{aligned} d(I_G, I_P) = {} & \alpha_{\mathrm{HEW}} \cdot d_{\mathrm{HEW}}(\mathrm{HEW}(I_G), \mathrm{HEW}(I_P)) + \alpha_{\mathrm{HSNS}} \cdot d_{\mathrm{HSNS}}(\mathrm{HSNS}(I_G), \mathrm{HSNS}(I_P)) \\ & + \alpha_{\mathrm{HTNS}} \cdot d_{\mathrm{HTNS}}(\mathrm{HTNS}(I_G), \mathrm{HTNS}(I_P)) + \alpha_{\mathrm{SF}} \cdot d_{\mathrm{SF}}(\mathrm{SF}(I_G), \mathrm{SF}(I_P)), \end{aligned} \tag{11}$$

where the $\mathrm{HEW}(\cdot)$, $\mathrm{HSNS}(\cdot)$, $\mathrm{HTNS}(\cdot)$ and $\mathrm{SF}(\cdot)$ are the histogram of the edge weight, the histogram of the spatial node strength, the histogram of the temporal node strength, and the skeleton features respectively. $d_\vartheta(.,.)$ measures the distance between two type of the ϑ features in gallery and probe sets and α_xs are normalized weights.

However, in the single-shot person re-identification case, the HTNS features are not used in matching. Thus matching distance d is defined as

$$\begin{aligned} d(I_G, I_P) = {} & \alpha_{\mathrm{HEW}} \cdot d_{\mathrm{HEW}}(\mathrm{HEW}(I_G), \mathrm{HEW}(I_P)) + \alpha_{\mathrm{HSNS}} \cdot d_{\mathrm{HSNS}}(\mathrm{HSNS}(I_G), \mathrm{HSNS}(I_P)) \\ & + \alpha_{\mathrm{SF}} \cdot d_{\mathrm{SF}}(\mathrm{SF}(I_G), \mathrm{SF}(I_P)). \end{aligned} \tag{12}$$

In our experiments, we fix the values of the parameters in the multi-shot case as: $\alpha_{\mathrm{HEW}} = 0.2$, $\alpha_{\mathrm{HSNS}} = 0.2$, $\alpha_{\mathrm{HTNS}} = 0.3$ and $\alpha_{\mathrm{SF}} = 0.3$, while in the single-shot case as: $\alpha_{\mathrm{HEW}} = 0.2$, $\alpha_{\mathrm{HSNS}} = 0.4$ and $\alpha_{\mathrm{SF}} = 0.4$. These values have been estimated once with cross-validation using a subset of 40 images of the KinectREID database. We use these weights in next stages.

Totally, there are three databases for person re-identification which were obtained by RGB-D sensors. One of them, the BIWI RGBD-ID database [27], was designed for long-term person re-identification and thus is not suitable for our work (i.e., short-term person re-identification). We perform our experiments on two databases, RGBD-ID database [25] and KinectREID database [28]. The RGBD-ID database contains RGB and depth data of 79 subjects. Four acquisitions were made for each subject, one rear and three frontal poses, and in one of the latter the arms were stretched; for each acquisition only 4 or 5 RGB-D frames were provided; sometimes, the same subject wore different clothes in different acquisitions: we removed the corresponding tracks. Hence, 2 to 4 acquisitions remained for each subject, for a total of 197 video sequences out of the 320 original ones. The KinectREID database includes video sequences of 71 subjects in three viewpoints: three near-frontal views, three near-rear views, and one lateral view. All the subjects walked normally; some of them carried accessories like bags. Seven video sequences were taken for each subject. The database includes RGB images, the segmentation masks, the skeletons, the 2D depth images, the 3D matrices of the depth images and the estimated floor.

We perform our experiments in the single-shot and multi-shot cases. For the former, one image of each person is randomly selected and forms the gallery set, and the rest images form the probe set. This procedure is repeated 10 times and computes average of the Cumulative Matching Characteristic (CMC) curves. The CMC curves treat re-identification as a ranking problem by representing the probability of finding the correct match over the first k ranks. In other words, $\mathrm{CMC}(k)$ can be seen as the recall at k. For the latter, both gallery and probe sets are made up of multi-shot signatures. The multi-shot signatures are built from N images of the same subject randomly selected. The single-shot case is labeled as SvsS and the multi-shot case is labeled as MvsM in this paper. We test our features for the MvsM case with $N = 5, 10$ on the KinectREID database and $N = 4/5$ on the RGBD-ID database (maximum the number of the frames in the RGBD-ID database is 4 or 5). For the MvsM case, we run 100 independent trials. In the evaluating, the parameter value of r, δ are selected $r = 3$ and $\delta = 30$ because these values have achieved better results than the other values that is proved by experiment in Subsection 4.3.

4.2 Comparison with other methods

We firstly compare our features with other features. So far most of the RGB-D re-identification approaches have used of only almost the same skeleton-based features such as [25, 29]. Since our method combines the appearance-based features with skeleton-based features and according to finding of [28], the skeleton

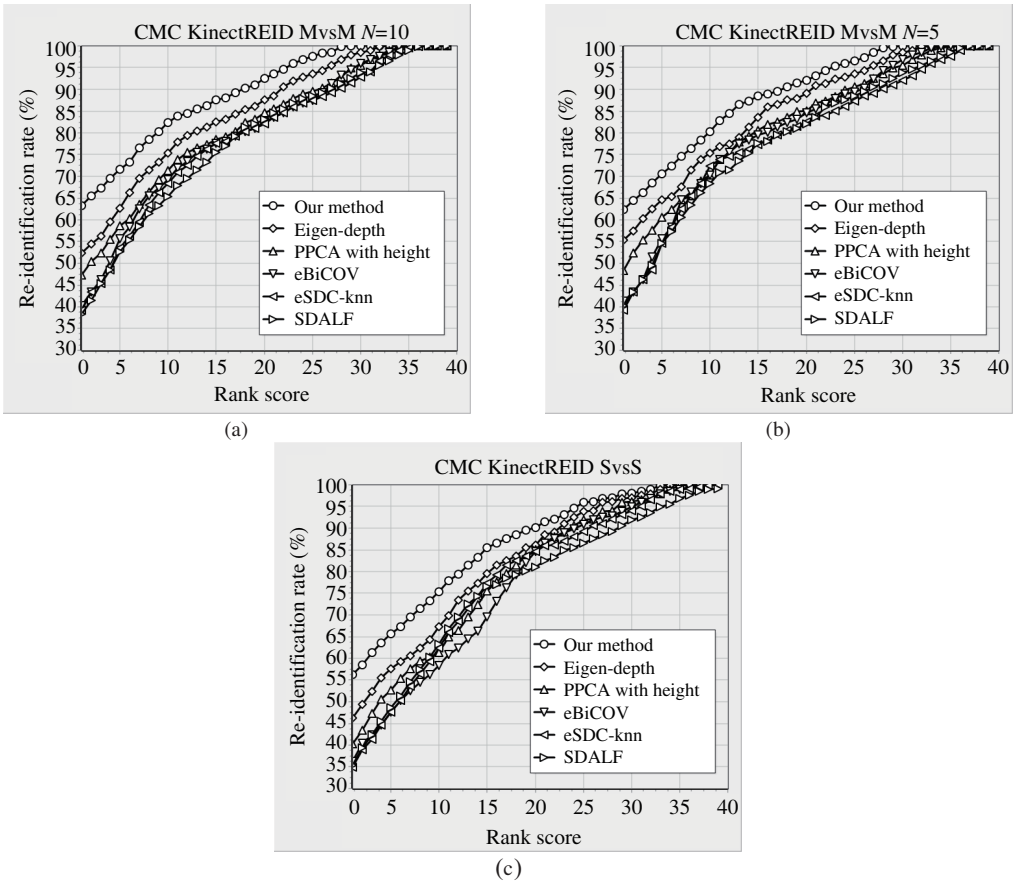

Figure 5 Performance comparison using the top 40 ranks of CMC curves on the KinectREID database.

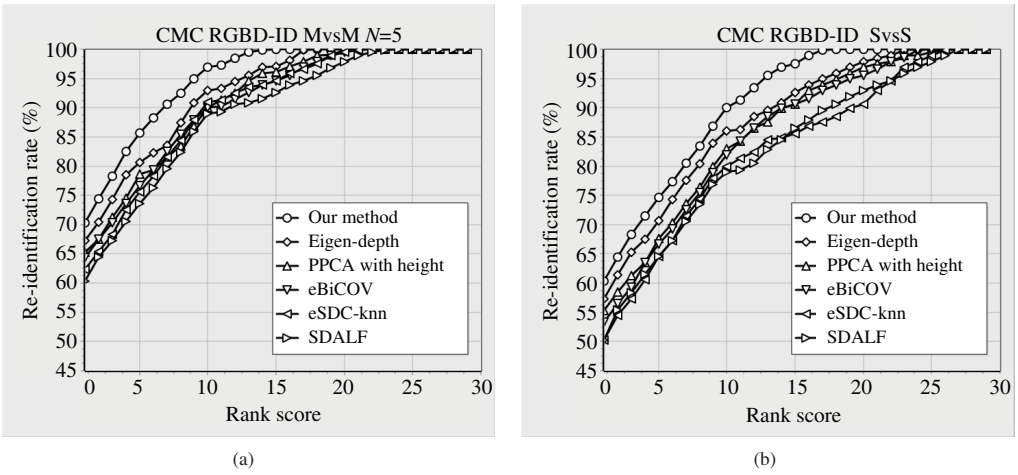

Figure 6 Performance comparison using the top 30 ranks of CMC curves on the RGBD-ID database.

features improves the re-identification accuracy based on the appearance features. Thus, we do not compare our features with these features. We compare our features with two the RGB-D appearance-based features of Eigen-depth [31], latent features [30] and also with SDALF [7], eBiCOV [8] and eSDC-knn [10] methods. For this reason, we combine all the methods with skeleton features using the score-level fusion. Figures 5 and 6 show CMC curves of the methods on the KinectREID and RGBD-ID databases respectively. It is seen from Figures 5 and 6 that there is obvious improvement compared with our features

Table 1 Comparison with other metrics using the Rank-1 and Rank-10 accuracies

Methods	Re-identification rate (%)			
	KinectREID		RGBD-ID	
	Rank-1	Rank-10	Rank-1	Rank-10
Score-level fusion (SvsS)	58.35	85.64	62.43	90.99
Score-level fusion (MvsM, $N = 5$)	63.21	86.35	72.35	97.99
Score-level fusion (MvsM, $N = 10$)	64.43	87.08	-	-
RankSVM (SvsS)	54.21	77.56	54.35	87.43
RankSVM (MvsM, $N = 5$)	56.21	82.43	55.46	89.72
RankSVM (MvsM, $N = 10$)	56.56	83.56	-	-
KISSME (SvsS)	57.35	84.35	61.08	87.64
KISSME (MvsM, $N = 5$)	59.21	85.71	63.24	90.46
KISSME (MvsM, $N = 10$)	60.35	87.86	-	-

against other features. For example, on KinectREID database in case of MvsM $N = 5$, rank-1 accuracy of our method, Eigen-depth, PPCA with height, eBiCOV, eSDC-knn and SDALF are of 62.35%, 55.35%, 48.21%, 41.35%, 39.64% and 40.43% respectively. Also rank-10 accuracy of our method, Eigen-depth, PPCA with height, eBiCOV, eSDC-knn and SDALF are of 80.35%, 75.56%, 70.35%, 70.35%, 72.08% and 68.56% respectively. Thus rank-1 accuracy is almost improved from 7% to 23%. Also, rank-10 accuracy is almost improved from 5% to 12%. As to RGBD-ID database in case of MvsM $N = 5$, rank-1 accuracy of our method, Eigen-depth, PPCA with height, eBiCOV, eSDC-knn and SDALF are of 70.35%, 67.21%, 65.21%, 64.56%, 62.5% and 60.43% respectively. Also rank-10 accuracy of our method, Eigen-depth, PPCA with height, eBiCOV, eSDC-knn and SDALF are of 96.99%, 92.86%, 89.35%, 89.78%, 90.78% and 88.56% respectively. Thus rank-1 accuracy is almost improved from 3% to 10%. Also, rank-10 accuracy is almost improved from 4% to 8%. As to other cases, results is almost consistent with case of MvsM $N = 5$.

We subsequently compare the score-level fusion method with other fusion methods. As to this case, we use the protocol [18] and divide databases to training and testing sets. Then randomly select $p = 30$ persons for training and model the methods using all possible positive and negative pairs in the training set. During testing, we divide test set to gallery and probe sets. Subsequently, the whole of methods are evaluated on testing set. For this comparison, the proposed features are employed with two metric learning algorithm namely RankSVM [14] and KISSME [19]. Table 1 shows the matching rates at rank-1 and rank-10 for the methods in the single-shot and the multi-shot cases. It is seen from Table 1 that only in a case of (MvsM, $N = 10$) on KinectREID database, Rank-10 of the KISSME is better than our method however totally ours outperforms the other methods both in the single-shot and the multi-shot cases.

For example, on KinectREID database in case of MvsM $N = 5$, rank-1 accuracy of the score-level fusion, RankSVM and KISSME are of 63.21%, 56.21% and 59.21% respectively. Also rank-10 accuracy of the score-level fusion, RankSVM and KISSME are of 86.35%, 82.43% and 85.71% respectively. Thus rank-1 accuracy is almost improved from 4% to 7%. Also, rank-10 accuracy is almost improved from 1% to 4%. As to RGBD-ID database in case of MvsM $N = 5$, rank-1 accuracy of the score-level fusion, RankSVM and KISSME are of 72.35%, 55.46% and 63.24% respectively. Also rank-10 accuracy of the score-level fusion, RankSVM and KISSME are of 97.99%, 89.72% and 90.46% respectively. Thus rank-1 accuracy is almost improved from 9% to 17%. Also, rank-10 accuracy is almost improved from 7% to 8%. As to other cases, results is almost consistent with case of MvsM $N = 5$. There is no publicly available code or the experimental results for majority of the existing methods on RGB-D databases. Thus we cannot perform more comparsions.

4.3 Parameter selection

There are two important parameters that might affect the performance critically. The first one is the neighborhood radius, r, and the second one is the threshold d. We evaluate the rank-1 accuracy on

Table 2 Rank-1 accuracies for different radiuses and thresholds on the KinectREID database

δ	Re-identification rate (%)				
	r				
	1	2	3	4	5
10	59.35	62.21	59.35	50.35	45.08
15	57.64	62.08	62.56	55.08	46.21
20	56.43	62.99	63.08	61.43	48.35
25	55.35	62.56	63.21	60.35	50.86
30	54.35	59.43	**63.35**	60.56	54.35
35	52.21	57.35	63.08	59.35	56.71
40	49.35	55.43	61.43	58.71	55.43

Table 3 Average computation time for different methods of extracting feature on KinectREID database

Method	The average time of the future extraction (ms)
Ours	97
Eigen depth	317
PPCA with height	413
eSDC_knn	192
eBiCOV	306
SDALF	251

the KinectREID database for different radiuses and thresholds. The rank-1 accuracies are illustrated in Table 2. From these results are seen that there arent regular procedure in the results improvement by increasing or decreasing parameters however best results are obtained in $r = 3$ and $\delta = 30$.

4.4 Computational performance

As to the efficiency of our approach, first, we compare the average computation time of different methods of extracting feature. The whole of experiments in this subsection are implemented on database KinectREID in multi-shot case ($N = 10$) by using an ASUS laptop with an Intel i5 2.7 GHZ processor and 6 GB of RAM, and software of Delphi 10.1. Processing times are averaged over ten runs of the experiments. The computation times of different features have been shown in Table 3. As can be seen, the time cost the extraction of latent feature is the most because it uses EM algorithm, a time-consuming algorithm, in feature extraction stages. The time cost of eBiCOV and Eg-depth features are almost close together and after latent features spend much more time because both the features use covariance descriptors for extracting feature which is time-consuming too. Also, eBiCOV combines the SDALF with gBiCOV so it is fully reasonable that time cost of the eBiCOV is more than the SDALF. In the other hand, the SDALF descriptor requires almost much time for parts segmentation and also number of features extracted in SDALF descriptor is more than eSDC-knn. Thus the time cost of the SDALF is more than eSDC-knn descriptor. It should be noted that in RGB-D databases, the background subtraction is performed using the sensors easily. This task requires almost much time in RGB databases so the running of the SDALF descriptor requires more time on RGB databases than the RGB-D databases. Finally, our method does not employ the heavy preprocessing steps as parts segmentation in the SDALF and does not use the time-consuming algorithms such as EM algorithm or covariance descriptors thus it is fully reasonable that the time cost of our method is the least.

Then we compare the average time computation of our matching approach, i.e., score-level fusion with the matching approaches of RankSVM and KISSME. Each of the three methods is employed on ourselves features. The time cost of the methods have been shown in Table 4. It is seen that RankSVM learning metric is the most time-consuming for re-identifying the persons against KISSME and score-level fusion, however the score-level fusion is the least time-consuming. It is fully reasonable because the score-level fusion only requires the computing of two distances derived by two types of features (features based on

Table 4 Average computation time for different methods of distance learning on KinectREID database

Method	The average time of the feature matching (ms)
Score-level fusion	43
RankSVM	503
KISSME	376

complex networks and skeleton features), while RankSVM and KISSME employ heavier computation steps than the score-level fusion. As expected, our method clearly outperforms the other methods in terms of the computational performance.

5 Conclusion

In this task, first, the histogram of the edge weight (HEW) and the histogram of the node strength (HNS) are extracted from graph built using the depth images. Then histograms are combined with the skeleton features using the score-level fusion and used for short-term person re-identification. While RGB sensors require complex calibration methods, and are very sensitive to occlusions, background clutter and lighting changes, the RGB-D sensors decrease the complexity of the person re-identification.

Our algorithm fits the single-shot and multi-shot person re-identification cases. In the single-shot case, the histogram of the node strength (HNS) is extracted from only one frame and called histogram of the spatial node strength (HSNS). In multi-shot case, in addition to extracting of the HSNS, the histogram of the temporal node strength (HTNS) is extracted from multi-frame.

Experiments on re-identification databases built using the Kinect (i.e., the KinectREID and RGBD-ID databases) show that the proposed method achieves adaptable performance and even better than the existing methods for person re-identification.

References

1 Guillaumin M, Verbeek J, Schmid C. Is that you? Metric learning approaches for face identification. In: Proceedings of 12th International Conference on Computer Vision (ICCV), Kyoto, 2009. 498–505
2 Wang L, Tan T N, Ning H Z, et al. Silhouette analysis based gait recognition for human identification. IEEE Trans Pattern Anal Mach Intell, 2003, 25: 1505–1518
3 Cong D N T, Khoudour L, Achard C, et al. People re-identification by spectral classification of silhouettes. Signal Process, 2010, 90: 2362–2374
4 Bedagkar-Gala A, Shah S K. A survey of approaches and trends in person re-identification. Image Vis Comput, 2014, 32: 270–286
5 Geng Y B, Hu H M, Zeng G D, et al. A person re-identification algorithm by exploiting region-based feature salience. J Vis Commun Image Represent, 2015, 29: 89–102
6 Schwartz W, Davis L. Learning discriminative appearance-based models using partial least squares. In: Proceedings of XXII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI), Rio de Janiero, 2009. 322–329
7 Farenzena M, Bazzani L, Perina A, et al. Person re-identification by symmetry-driven accumulation of local features. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, 2010. 2360–2367
8 Ma B P, Su Y, Jurie F. Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image Vis Comput, 2014, 32: 379–390
9 Cheng D S, Cristani M, Stoppa M, et al. Custom pictorial structures for re-identification. In: Proceedings of the British Machine Vision Conference (BMVC), Dundee, 2011. 68.1–68.11
10 Zhao R, Ouyang W L, Wang X G. Unsupervised salience learning for person re-identification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, 2013. 3586–3593
11 Zhao R, Oyang W L, Wang X G. Person re-identification by saliency learning. IEEE Trans Pattern Anal Mach Intell, 2017, 39: 356–370
12 Si J L, Zhang H G, Li C-G, et al. Spatial pyramid-based statistical features for person re-identification: a comprehensive evaluation. IEEE Trans Syst Man Cybern Syst, 2017, 99: 1–5
13 Lan R S, Zhou Y C, Tang Y Y. Quaternionic local ranking binary pattern: a local descriptor of color images. IEEE Trans Image Process, 2016, 25: 566–579
14 Prosser B, Zheng W S, Gong S G, et al. Person re-identification by support vector ranking. In: Proceedings of the British Machine Vision Conference (BMVC), Aberystwyth, 2010. 21.1–21.11

15 Bashir K, Xiang T, Gong S G. Feature selection on gait energy image for human identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, 2008. 985–988

16 Zheng W S, Gong S G, Xiang T. Person re-identification by probabilistic relative distance comparison. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, 2011. 649–656

17 Loy C C, Liu C X, Gong S G. Person re-identification by manifold ranking. In: Proceedings of IEEE International Conference on Image Processing, Melbourne, 2013. 3567–3571

18 Figueira D, Bazzani L, Minh H Q, et al. Semisupervised multi-feature learning for person re-identification. In: Proceedings of 10th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Krakow, 2013. 111–116

19 Kostinger M, Hirzer M, Wohlhart P, et al. Large scale metric learning from equivalence constraints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, 2012. 2288–2295

20 Tao D P, Guo Y N, Song M L, et al. Person re-identification by dual-regularized KISS metric learning. IEEE Trans Image Process, 2016, 25: 2726–2738

21 Mignon A, Jurie F. PCCA: a new approach for distance learning from sparse pairwise constraints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, 2012. 2666–2672

22 Liu H, Ma B P, Qin L, et al. Set-label modeling and deep metric learning on person re-identification. Neurocomput, 2015, 151: 1283–1292

23 Vezzani R, Baltieri D, Cucchiara R. People re-identification in surveillance and forensics: a survey. ACM Comput Surv, 2013, 46: 29

24 Chen S-Z, Guo C-C, Lai J-H. Deep ranking for person re-identification via joint representation learning. IEEE Trans Image Process, 2016, 25: 2353–2367

25 Barbosa I B, Cristani M, Bue A D, et al. Re-identification with RGB-D sensors. In: Fusiello A, Murino V, Cucchiara R, eds. Computer Vision ECCV, Workshops and Demonstrations. Berlin/Heidelberg: Springer-Verlag, 2012. 433–442

26 Satta R, Pala F, Fumera G, et al. Real-time appearance-based person re-identification over multiple kinect cameras. In: Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP), Barcelona, 2013. 21–24

27 Munaro M, Fossati A, Basso A, et al. One-shot person re-identification with a consumer depth camera. In: Gong S G, Cristani M, Yan S C, et al., eds. Person Re-identification. London: Springer, 2014. 161–181

28 Pala F, Satta R, Fumera G, et al. Multi-modal person re-identification using RGB-D cameras. IEEE Trans Circuit Syst Video Technol, 2015, 26: 788–799

29 Munaro M B, Ghidoni S, Tartaro D T, et al. A feature-based approach to people re-identification using skeleton keypoints. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, 2014. 5644–5651

30 Oliver J, Albiol A, Albiol A, et al. Using latent features for short-term person re-identification with RGB-D cameras. Pattern Anal Appl, 2015, 19: 549–561

31 Wu A C, Zheng W S, Lai J H. Depth-based person re-identification. In: Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition, Kuala Lumpur, 2015. 026–030

32 Backes A R, Casanova D, Bruno O M. Texture analysis and classification: a complex network-based approach. Inf Sci, 2013, 219: 168–180

33 Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization. Nat Rev Genet, 2004, 5: 101–113

34 Newman M E J, Park J. Why social networks are different from other types of networks. Phys Rev E, 2003, 68: 036122

35 Gonçalves W, Martinez A, Bruno O. Complex network classification using partially self-avoiding deterministic walks. Chaos, 2012, 22: 033139

36 Gonçalves W N, Machado B B, Bruno O M. A complex network approach for dynamic texture recognition. Neurocomput, 2015, 153: 211–220

37 Amaral L A N, Ottino J M. Complex networks: augmenting the framework for the study of complex systems. Eur Phys J B, 2004, 38: 147–162

38 Gonçalves W N, Backes A R, Martinez A S, et al. Texture descriptor based on partially self-avoiding deterministic walker on networks. Expert Syst Appl, 2012, 39: 11818–11829

39 Gonçalves W N, Silva J A, Bruno O M. A rotation invariant face recognition method based on complex network. In: Bloch I, Cesar R M, eds. Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2010. Berlin/Heidelberg: Springer, 2010. 426–433

40 Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: a survey of measurements. Adv Phys, 2006, 56: 167–242

41 Bazzani L, Farenzena M, Perina A, et al. Multiple-shot person re-identification by HPE signature. In: Proceedings of IEEE International Conference on Pattern Recognition, Istanbul, 2010. 1413–1416