Scientia Iranica (2024) 31(2), 103–117

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
https://scientiairanica.sharif.edu

Re-identification in video surveillance systems considering appearance changes

Z. Mortezaie[a], **H. Hassanpour**[a,*], **and A. Beghdadi**[b]

a. *Faculty of Computer Engineering, shahrood University of Technology, Shahrood, Iran.*
b. *Institut Galilée, Université Sorbonne Paris Nord, Villetaneuse, France.*

Received 24 January 2021; received in revised form 29 May 2022; accepted 10 September 2023

KEYWORDS
Surveillance systems;
Re-identification;
Appearance changes;
Robust descriptor;
Weight map.

Abstract. Human behavior analysis and visual anomaly detection are important applications in fields such as video surveillance, security systems, intelligent houses, and elderly care. People re-identification is one of the main steps in a surveillance system that directly affects system performance; and variations in appearance, pose, and scene illumination may be challenging issues for such a system. Previous re-identification approaches faced limitations while considering appearance changes in their tracking task. This paper proposes a new approach for people's re-identification using a descriptor that is robust to appearance changes. In our proposed method, the enhanced Gaussian Of Gaussian (GOG) and the Hierarchical Gaussian Descriptors (HGDs) are employed to extract feature vectors from images. Experimental results on a number of commonly used people re-identification databases imply the superiority of the proposed approach in people re-identification compared to other existing approaches.

1. Introduction

Video surveillance systems have become an indispensable appliance of modern safety management in both public and private sectors. Surveillance systems using network cameras usually employ cameras with non-overlapping fields of view to cover a broader area with fewer cameras. People re-identification is an important issue in surveillance systems. Indeed, a higher accuracy in people re-identification leads to performance improvements in surveillance systems. People re-identification is defined as observing a person in a camera and re-identifying the person in the same or subsequent cameras in the network after a while. Indeed, re-identification systems are similar to identification systems, which include a set of images associated with the identified people, namely the *gallery set*, and an image of an unidentified or newly arrived person in a camera view, which is a member of the *probe set.*

Note that in real situations, the gallery set and probe set may change over time. However, for simplicity, the number of members in the gallery set and probe set is considered fixed in a given time interval. Suppose that the gallery set for a given system is as follows:

$$G = \{g_1, g_2, \cdots, g_N\}, \tag{1}$$

where g_i, $i = 1, \cdots, N$, is the identified person. The assigned labels to the identified persons are as follows:

$$id_G = \{(g_1, id(g_1)), (g_2, id(g_2)), ..., (g_N, id(g_N))\}, \tag{2}$$

where $id(g_i)$ determines the label associated with the

[*] *Corresponding author.*
E-mail addresses: zm.mortezaie@gmail.com (Z. Mortezaie); h_hassanpour@yahoo.com and h.hassanpour@shahroodut.ac.ir (H. Hassanpour); beghdadi@sorbonne-paris-nord.fr (A. Beghdadi)

doi: 10.24200/sci.2023.57617.5331

gallery image g_i. Also, suppose that the probe set is defined as:

$$P = \{p_1, p_2, \cdots, p_M\}, \tag{3}$$

In a re-identification system, the goal is to determine the labels of probe set members. This goal is achieved by matching each un-identified person p with all of the identified people as follows:

$$id(p) = \begin{cases} id(g_i) \quad arg\max_i \text{ (similarity } (p, g_i) > T), \\ \qquad i = 1, 2, \cdots, N \\ \beta \qquad \text{otherwise} \end{cases} \tag{4}$$

Indeed, the similarity between each member of the probe set and members of the gallery set is measured. If the similarity is higher than a threshold (T), the label of the most similar member of the gallery set is assigned to the unidentified member. Otherwise, the unidentified member is considered a newly arrived person in the camera view. Hence, a new label (β) is assigned to the person, and the gallery set is updated.

Many approaches have attempted to improve the performance of the re-identification task, some of which use a person's appearance for re-identification. Other approaches use additional distinctive characteristics such as a person carrying objects or the color of their clothing. Many of the existing approaches assume that the subject's appearance doesn't change during the re-identification task. This assumption is one of the limitations of existing re-identification systems, which reduces their flexibility. In addition, changing pose in the camera's field of view or the absence of a previously carried object may disrupt the distinctive characteristics of the person. In these situations, saliency-based approaches do not perform well for people re-identification.

An approach is proposed in this paper for people re-identification using a descriptor that is robust to appearance changes. The contribution of our proposed method is to enhance the commonly used descriptors in people re-identification, namely Gaussian Of Gaussian (GOG) [1] and the Hierarchical Gaussian Descriptor (HGD), which is an extended version of GOG [2]. The performance of these descriptors is enhanced using a weighing mechanism.

In the proposed weighing mechanism, the effect of each pixel on the extracted features is considered proportional to its association with the background and person's body. Our proposed weighing mechanism can be considered as an unsupervised semantic segmentation approach, where it assigns pixels with the same importance to the same segment without requiring the training phase and labeled samples.

Our proposed approach is evaluated using several common people re-identification databases. Experimental results denote the superiority of our proposed approach compared to other existing approaches for people re-identification.

This paper is structured as follows: the existing people re-identification approaches are briefly reviewed in Section 2. The proposed re-identification method is introduced in Section 3. The experimental results and conclusion are presented in Sections 4 and 5, respectively.

2. Related works

Many approaches have been proposed to improve the performance of re-identification systems. The existing re-identification approaches can be generally grouped into two main categories: methods that only use the appearance of people and those that use both the appearance and salient features extracted from detected people. Methods in the first category mainly focus on the color or texture of the person's clothes, while approaches in the second category focus on distinctive characteristics such as carrying special objects or wearing clothes with a specific color. In this section, we briefly review some of the existing re-identification approaches belonging to the above-mentioned categories.

2.1. Approaches based on appearance

In the appearance-based approaches, features such as color and texture of clothes are used for re-identification. In [3], the local details of a person's appearance are described using local patches with a 50% overlap. In this method, two scales of Scale Invariant Local Ternary Pattern (SILTP) [4] histograms and an $8 \times 8 \times 8$-bin joint HSV histogram are extracted from each patch, where each bin describes the occurrence probability of the corresponding pattern in the patch. Then, the local occurrence of each pattern (i.e., the same histogram bin) is maximized by analyzing the patches at the same horizontal location in order to obtain features robust to viewpoint changes. Also, in this method, a subspace and the cross-view Quadratic Discriminant Analysis (XQDA) method were proposed to learn a discriminant low-dimensional subspace. Meanwhile, in this approach, a quadratic discriminant analysis metric is learned on the derived subspace simultaneously.

In the method proposed in [5], a Hierarchical Feature Model (HFM) has been proposed for multi-target tracking in video frames. In this method, using extracted features from the lower layers of GoogLeNet, a dictionary of appearance characteristics is learned. Then, Orthogonal Matching Pursuit (OMP) [6] is applied to the extracted features to obtain a sparse representation of the feature vectors, which was used as a dictionary. Finally, people tracking is performed using a Bayesian filter and a discrete combinatoril optimization.

Also, in [7], based on Convolutional Neural Networks (CNNs), a parallel spatial-temporal attention model is proposed to re-identify people in video sequences. The goal of this model is to extract temporal and spatial features simultaneously without losing space information.

Note that CNN-based methods face two major drawbacks. First, they need to be trained on very large datasets to converge on the training objective, and second, training CNN models is very time consuming due to the very large number of -trainable parameters, which all need to be updated on every training step.

As mentioned before, one of the straight assumptions in existing people re-identification approaches is that the target's appearance doesn't change during the tracking process. However, in real situations, the target's appearance may vary across cameras in a surveillance system.

In [8], a depth-based person re-identification model was proposed to overcome the problem of clothes changing and extreme illumination. In this method, a depth-voxel covariance descriptor (within-voxel covariance and between-voxel covariance) and a locally rotation invariant depth shape descriptor named Eigen-depth feature have been proposed to describe the person's body shape.

In addition, to cover some existing issues in re-identification systems, such as changes in illumination and orientation, an appearance-based descriptor has been proposed in [1]. In this method, namely GOG, a local region in an image is described via hierarchical Gaussian distribution in which both means and covariance are included in its parameters. In this method, the images are divided into seven overlapping horizontal stripes. Then, each region is further divided into a number of patches in order to present the local structure of each region. For describing each pixel in the patches, a feature vector is used, which involves some raw features such as pixel location in the vertical direction, the magnitudes of the pixel intensity gradient in four different orientations, and the color channel values in R, G, and B components. In this method, each region is modeled as a set of multiple Gaussian distributions using the Gaussian distributions of its patches. Meanwhile, the effect of each patch is proportional to its distance to the center of the image. Indeed, assuming the person's body is located at the center of the image, the closer the patch is to the center of the image, the higher its effect is on the Gaussian distribution. As the Gaussian distribution is non-linear, GOG's the parameters of both patches and region distributions are mapped into the tangent space. Finally, the mapped Gaussian distributions of the regions are used to describe the whole input image.

The HGD was proposed in [2] as an enhanced version of GOG. In this approach, feature norm normalization methods are developed in order to reduce the bias of Symmetric Positive Definite (SPD) matrix descriptors.

The same feature extraction manner as GOG is used in another descriptor, namely the Multi-Level Gaussian Descriptor (MLGD) proposed in [9]. In this approach, features such as color moment values of RGB components and Schmid filter responses for representing each pixel of the image make it slightly different from GOG.

In [10], a Kernel cross-view Collaborative Repre -sentation-based Classification (Kernel X-CRC) approach is proposed to handle the issue of appearance changes caused by different camera conditions. This approach uses the GOG descriptor for extracting images' appearance features, which are further mapped into the learned subspaces. The mapped features are passed through the Kernel X-CRC to compute the similarity between images.

To cover the issues of variations in viewpoint and pose, a Graph Correspondence Transfer (GCT) method was proposed in [11]. In this approach, first, a patch-wise graph matching mechanism is used for learning a set of patch-wise correspondence templates from positive image pairs with various pose-pair configurations. Then, for each pair of test images, some training pairs with the most similar pose-pair configurations are selected as references. Then, to compute the similarity between images, the correspondences of the references are transferred to the test pair. Also, a pose context descriptor based on the topology structure of the estimated joint locations [12] is used in [13] to empower the correspondence transfer used in [11].

In [14], a Sample-Specific Multi-Kernel (SSMK) approach was proposed to handle the issue of appearance changes across camera views. In this approach, the images are horizontally divided into six regions. Also, some features such as RGB, YUV, HSV, LAB, and YCbCr color information, Dense SIFT [15,16], color naming feature [17], and deep features [18] are extracted from each region. The extracted features are mapped into the weighed multi-kernel feature space. The mapped features are further used to learn a discriminative metric for re-identification.

In [19], for reducing the number of labeled training samples in person re-identification, a View-Specific Semi-supervised Subspace Learning (VS-SSL) approach is used, where specific projections are learned for each camera view. In this approach, the appearance features extracted from GOG are used.

In [20], people re-identification is considered a consistent iterative multi-view joint transfer learning optimal problem, where the Inexact Augmented Lagrange Multiplier (IALM) algorithm [21] is used to solve the problem. The goal of this approach is to cover the issue of data distribution inconsistency between camera views.

In [22], SVDNet is proposed to extract global appearance features from CNN by optimizing the deep representation learning process using the Singular Vector Decomposition (SVD).

To handle the issue of misaligned images, based on the CNN feature maps, the Pedestrian Alignment Network (PAN) is proposed in [23], where it aligns the pedestrians within bounding boxes and learns the pedestrian descriptors simultaneously.

In [24], to re-identify each probe image, a number of investigators are simulated by combining some appearance-based feature extraction approaches with various metric learning methods in pairs, where the output of each pair is a ranking list for the probe image. Then, the obtained ranking lists are fused using a proposed crowdsourcing-based ranking aggregation approach. In this re-identification approach, the appearance features extraction approaches introduced in [1,25,26] are used for representing a person's images.

In [27], a Multi-level Feature Fusion (MFF) approach is proposed to extract discriminative appearance characteristics, where the global and local features of a person's images are fused through deep learning networks. In this approach, to fuse low-to-high-level local features, the local features are extracted from different layers of the deep network. Also, Global-Local Branches are used to extract the local and global features at the highest level.

Also, Leng [28] proposed a semi-supervised co-metric learning approach, where a few annotated training samples are used to train a discriminative Mahalanobis-like distance matrix. This approach, first, uses both the hand-crafted features, i.e., color name [17], and the features extracted from the Siamese CNN [29] for representing single-view person images. Meanwhile, to decompose the single-view person features into pseudo-binary views, a binary-weight learning approach is used. The decomposed features are then used to train metric models, where the metric models are updated jointly using both the pseudo labels and references to obtain discriminative metrics.

Also, in [30], the HSV and SILTP features are extracted in the local form to handle the issue of occlusion and pose variations in people's re-identification. The approach proposed in [31] combined the features used in [3], as well as the extracted features from CNN, in order to improve the performance of people re-identification systems.

2.2. *Approaches based on both appearance and saliency*

In addition to changes in appearance, it is possible for a person to have distinctive characteristics. Hence, some approaches consider the distinctive characteristics of a weighing mechanism. In the method proposed in [32], first, a limited number of mid-level characteristics such as backpack, carrying object, short-hair, and long-hair are chosen by a human expert. Then, for each characteristic, a classifier is trained using a dataset. Then, for each unlabeled person, the existence of the considered mid-level characteristics is determined using the trained classifiers. In this method, the number of considered distinctive characteristics is limited. However, the distinctive characteristics are not predictable.

The saliency learning and matching framework, which was proposed in [33,34], can cover various distinctive characteristics. In this method, for each image in the gallery and probe sets, a saliency map is computed by comparing the patches of images with a specified reference set without any distinctive characteristics. Then, a saliency score for each image patch is estimated using K-Nearest Neighbors and one-class SVM [35]. Finally, re-identification is done by comparing the probe image saliency map with saliency maps of gallery images.

In [36], a kernelized graph-based approach was proposed for estimating the salient regions of a person's appearance as a saliency map. In this method, the saliency of each pixel of the input image is estimated considering its adjacent pixels. In addition, in this method, similar to [37], the Markov chain approach is used for the saliency map. Then, in the feature extraction step, each pixel of the obtained saliency map is used as a weight for the corresponding pixel in the input image. Hence, in the feature extraction step, color, shape, and texture characteristics are extracted both in weighted form and in non-weighted form. Then, the extracted features are used to learn a pairwise-based multiple metric. In this step, a non-Euclidean metric is learned for each feature. Finally, the learned metrics are combined for person re-identification.

In [38], a Harmonious Attention CNN (HA-CNN) module is proposed for jointly learning multi-granularity attention selection and feature representation. By using this approach, the correlated complementary information between attention selection and feature discrimination is maximized.

In the method proposed in [39], the regions of a person are extracted using mean shift [40]. Then, the global contrast-based salient region detection [41] is used to compute the saliency of each region. The salient regions are then clustered using least-squares log-density gradient clustering [42]. Hence, a cluster is associated with each person in the video frames. Finally, the distances between the salient region of the probe image and the clustered salient regions of the gallery images are determined for re-identification.

As mentioned before, the distinctive characteristics of a person may disappear for reasons such as a change in orientation or carrying/not carrying certain objects. In addition, assuming no change in appearance over time is another limitation of existing re-identification systems. The mentioned limitations in people's re-identification systems also reduce their performance. Hence, a system is proposed in this paper for people's re-identification using a descriptor robust to appearance changes.

3. Proposed method

As mentioned before, a re-identification system may encounter a person with various poses or orientations in different cameras. These changes may cause the system to miss the person during the tracking task. Hence, descriptors that are robust to appearance changes are required to successfully re-identify people in network cameras.

Robustness to variations in appearance can be achieved using appropriate discriminant descriptors such as the GOG [1] and HGD [2]. In Sub-section 3.1, we briefly review the classic GOG and HGD descriptors. Also, Sub-section 3.2 shows how the performance of these descriptors can improve.

3.1. Classic GOG and HGD descriptors

As mentioned before, in GOG and HGD descriptors, the image is first divided into seven overlapping horizontal stripes. Meanwhile, to depict the local structure of each region (strip), regions are divided into a number of patches. Also, a feature vector F_i is used to represent each pixel i in the patches. F_i involves a number of raw features such as pixel location in the vertical direction v in the image, the magnitudes of the pixel intensity gradient in four different orientations: D_{0°, D_{90°, D_{180°, D_{270°; and the color channel values in R, G, and B components; x_R, x_G, and x_B are obtained as follow:

$$F_i = [v; D_{0^\circ}; D_{90^\circ}; D_{180^\circ}; D_{270^\circ}; x_R; x_G; x_B]^T. \quad (5)$$

The patches and regions are further described using the Gaussian distributions in the following steps:

Step 1. For each patch s with l pixels, the corresponding raw features (i.e., $f = \{F_1, F_2, \cdots, F_l\}$) are used to compute a Gaussian distribution as follows:

$$\mathcal{N}(f; \mu_s, \Sigma_s) = \frac{\exp\left(-\frac{1}{2}(f-\mu_s)^T \Sigma_s^{-1}(f-\mu_s)\right)}{(2\pi)^{\frac{d}{2}} |\Sigma_s|}, \quad (6)$$

where μ_s and Σ_s respectively denote the mean and covariance matrix of the feature vectors associated with patch s, and d represents the size of the feature vector. Also, $|.|$ is the determinant operator.

Note that the Gaussian distribution space resides in a Riemannian manifold, and the Euclidean operation cannot be directly used in this manifold [43]. Besides, one of the Riemannian manifolds is SPD [44]. In SPD, the Log Euclidean Riemannian Metric (LERM) [45] can be used to map a point to Euclidean tangent space. Indeed, by using a principle matrix log, the SPD points can be mapped to the tangent space. Meanwhile, following the work in [46], the space of d-dimensional multivariate Gaussians can be placed in the space of $(d+1)$-dimensional SPD matrices. Hence, in GOG and HGD, first, the mean vector and covariance matrix of each patch are embedded into SPD using Eq. (7):

$$P_s = |\Sigma_s|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma_s + \mu_s \mu_s^T & \mu_s \\ \mu_s^T & 1 \end{bmatrix}. \quad (7)$$

The matrix P_s in Eq. (7) is then mapped to the tangent space using matrix logarithm as follows:

$$\begin{aligned} g_s &= vec(log(P_s)) \\ &= \big[\hat{p}_{s(1,1)}, \sqrt{2}\hat{p}_{s(1,2)}, \ldots, \sqrt{2}\hat{p}_{s(1,d+1)}, \\ &\quad \hat{p}_{s(2,2)}, \sqrt{2}\hat{p}_{s(2,3)}, \ldots, \hat{p}_{s(d+1,d+1)}\big]^T, \end{aligned} \quad (8)$$

where g_s is a vector with $b = \frac{(d^2+3d)}{2} + 1$ elements, and $log(.)$ denotes the matrix logarithm operator. Meanwhile, $log(P_s)$ is a symmetric matrix. Hence, in Eq. (8), the upper triangular part of the mapped matrix (i.e., $log(P_s)$) is obtained as a vector using the $vec(.)$ operator. Note that in this equation, $\hat{p}_{s(q,r)}$ is the element (q,r) of the $log(P_s)$ matrix.

Step 2. Each region $\mathcal{G}$ is modeled as a set of multiple Gaussian distributions using the Gaussian distributions of its patches as follows:

$$\mu^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s g_s, \quad (9)$$

$$\Sigma^{\mathcal{G}} = \frac{1}{\sum_{s\in\mathcal{G}} w_s} \sum_{s\in\mathcal{G}} w_s \left(g_s - \mu^{\mathcal{G}}\right)\left(g_s - \mu^{\mathcal{G}}\right)^T, \quad (10)$$

where $\mu^{\mathcal{G}}$ and $\Sigma^{\mathcal{G}}$ are the b dimensional mean vector and $b \times b$ dimensional covariance matrix of the region. Also, assuming that the person's body is located in the center of the image, w_s is used for tuning the influence of the patches' Gaussian distributions as follows:

$$w_s = \exp\left(\frac{-\frac{1}{2}(x_s - x_c)^2}{\sigma^2}\right), \quad (11)$$

where $x_c = \frac{w}{2}$, $\sigma = \frac{w}{4}$, and x_s show x-coordinates of the central pixel in patch s, and w is the width of the image.

Finally, the mean vector and covariance matrix of each region are mapped to the tangent space, as was done in the first step. Hence, each region $\mathcal{G}$ is described using a vector, namely $z_{\mathcal{G}}$ with $h = \frac{(b^2+3b)}{2} + 1$ elements. For each input image, the output of the GOG and HGD is a feature vector by concatenating the $z_{\mathcal{G}}$ associated with each of its seven regions as follows:

$$Z_{RGB} = \left[z_{\mathcal{G}1}^T, z_{\mathcal{G}2}^T, z_{\mathcal{G}3}^T, z_{\mathcal{G}4}^T, z_{\mathcal{G}5}^T, z_{\mathcal{G}6}^T, z_{\mathcal{G}7}^T\right]^T. \tag{12}$$

Also, due to the importance of color information in appearance-based person re-identification in GOG and HGD, the RGB information (i.e., x_R, x_G, and x_B) used in Eq. (5) is further substituted with the components of LAB, HSV, and components (nR) and (nG) of the nRGB (Normalized RGB i.e., $nR = \frac{R}{(R+G+B)}; nG = \frac{G}{(R+G+B)}; nB = \frac{B}{(R+G+B)}$) color spaces. By these substitutions, four different feature vectors are obtained for each image. Assume that the extracted feature vectors using RGB, LAB, HSV, and nRGB are named Z_{RGB}, Z_{LAB}, Z_{HSV}, and Z_{nRGB}, respectively. The final feature vector, which can be used to represent the input image, is obtained by concatenating Z_{RGB}, Z_{LAB}, Z_{HSV}, and Z_{nRGB} as follows:

$$Z_{Fusion} = \left[Z_{RGB}^T, Z_{LAB}^T, Z_{HSV}^T, Z_{nRG}^T\right]^T. \tag{13}$$

Note that for RGB, LAB, and HSV color spaces, the dimension of the feature vector F introduced in Eq. (5), i.e., d is 8, whereas d is 7 by considering components (nR) and (nG) of nRGB color space. Accordingly, for RGB, LAB, and HSV color spaces, the number of elements of vector g_s, i.e., b, is:

$$\frac{(8^2 + 3 \times 8)}{2} + 1 = 45,$$

and the number of elements of vector $z_{\mathcal{G}}$, i.e., h, is:

$$\frac{(45^2 + 3 \times 45)}{2} + 1 = 1081,$$

where as using two components of nRGB color space, b and h are as follows, respectively:

$$\frac{(7^2 + 3 \times 7)}{2} + 1 = 36,$$

$$\frac{(36^2 + 3 \times 36)}{2} + 1 = 703.$$

Consequently, considering seven regions in each image, the dimension of feature vector Z_{Fusion} is $3 \times 7 \times 1081 + 1 \times 7 \times 703 = 27622$.

According to Eq. (11), in GOG and HGD, the inverse of the image patche's distance from the image center has a straight effect on the regions' Gaussian distribution. Indeed, if a person is not located in the center of the image, these descriptors will associate a high weight to the background pixels. Consequently, the background pixels are mistakenly treated as parts of the person's body. Hence, it is necessary to propose an appropriate weighing mechanism that tunes the effect of each pixel in the final extracted features proportional to its association with the background or person's body.

In [1,2], the extracted feature vectors, i.e., Z_{Fusion}, are then used to train the XQDA distance metric, where, it simultaneously learns both a discriminative subspace and a distance metric by extending Bayesian face [47] and KISSME [48] approaches. This problem in XQDA is considered as a Generalized Rayleigh Quotient [49]. Besides, the generalized eigenvalue decomposition is used to achieve a closed- form solution. The output of XQDA is a score matrix. As mentioned before, each element of the score matrix shows the distance between the corresponding images in probe and gallery sets. Hence, by sorting each row of the score matrix, gallery images are ranked proportional to their distance to the corresponding probe image.

The details of the proposed re-identification approach are described in the next subsection. In this sub-section, we describe how the performance of the GOG descriptor can be improved using the proposed weighing approach. Then, our proposed approach is generalized to enhance the performance of HGD.

3.2. Enhancing the GOG descriptor

As mentioned before, in the GOG descriptor, the weighing mechanism is only appropriate when pixels in the middle of the image form the person's body. However, the person may not be located in the middle of the image in practice. In Figure 1(a), a sample image is shown. In this image, the person is not located in the middle of the image. The weight map used in the GOG descriptor for weighing the image patches is shown in Figure 1(b). As shown in this figure, the patch weighing mechanism used in the GOG descriptor

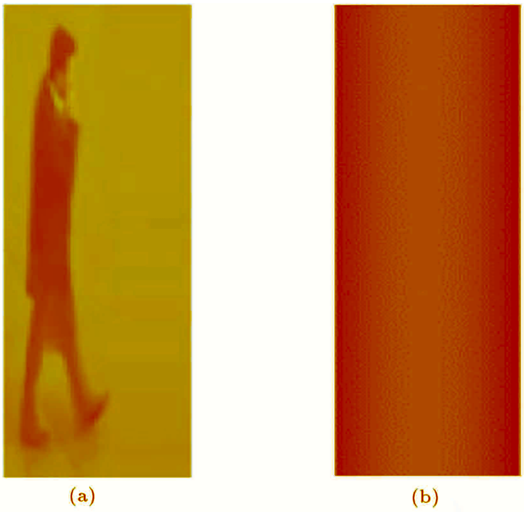

Figure 1. (a) A sample image and (b) the weight map used in [1].

focuses more on the background pixels than the pixels associated with the person's body.

The effect of each pixel in the final extracted features should be tuned proportional to its association with the background or the subject's body to enhance the GOG descriptor. Hence, in our proposed approach, for an input image, we compute a weight map with the same size as the image. Meanwhile, pixels of the weight map represent the effect of the corresponding input image pixels on the final extracted features.

In our proposed method, the weight map can be treated as a segmented image where the pixels with the same importance are assigned to the same segment. One of the commonly used color spaces for color image segmentation is the LAB [50]. LAB is a perceptually uniform color space [51], where the difference between color values is proportional to the difference of the comprehended colors. This capacity of LAB makes it a suitable color space for segmenting the color image. Hence, in the proposed approach, the LAB color space is used to obtain the weight map.

In real situations, the color of the subject's body is usually different from the background color. Besides, natural scene statistics of images can be described using Gaussian distributions [52]. Assume that the pixels of an input image are modeled using a Gaussian distribution. It is clear that the frequency of each pixel value is reflected in the obtained Gaussian distribution. In the input image, for example, if the number of pixels with a low value is more than the number of pixels with a high value, the mean of the Gaussian distribution tends to a low value. Hence, the probability that a pixel with a low value belongs to the distribution is high, and vice versa. Hence, assuming a difference between the values of background and subject body pixels, in our proposed approach, we model the image pixels using Gaussian distributions.

Assuming that the size of the input image is $m \times n \times 3$, the proposed weighing approach consists of three steps as follows:

Step 1. In the first step, the input RGB image is converted to the LAB color model. Then, we model the converted image (i.e., $image_{LAB}$) as a Gaussian distribution considering its L, A, and B components in the form of a $1 \times \mathcal{K}$ vector (i.e., $vector_L$, $vector_G$, and $vector_B$). Note that $\mathcal{K}$ is the number of pixels in the three components of the image, i.e., $\mathcal{K}=m \times n \times 3$.

Step 2. In the second step, for each pixel of the image, the Probability Density Function (PDF) is computed considering the mean and variance of the obtained distribution from the first step. Hence, for each pixel of the input image component, one value is obtained, which denotes the probability of that pixel belonging to the distribution. Hence, the outputs of the second step (i.e., PDF_L, PDF_A, and PDF_B) can be considered as an image with the same size as the input image (i.e., $m \times n \times 3$). We name this image W_{LAB}. The pixels in the image components (i.e., W_L, W_A, and W_B) represent the probabilities of the corresponding pixels in L, A, or B belonging to the input image Gaussian distribution.

Step 3. In the third step, W_{LAB} is converted to a normalized gray image with the size of $m \times n$ as follows:

$$\hat{W} = \frac{W_L + W_A + W_B}{\max(W_L + W_A + W_B)}. \tag{14}$$

In Figure 2, the main steps used for obtaining the proposed weight map ($\hat{W}$) are illustrated.

In Figure 3, a sample input image and the corresponding W_{LAB} and $\hat{W}$ images are shown.

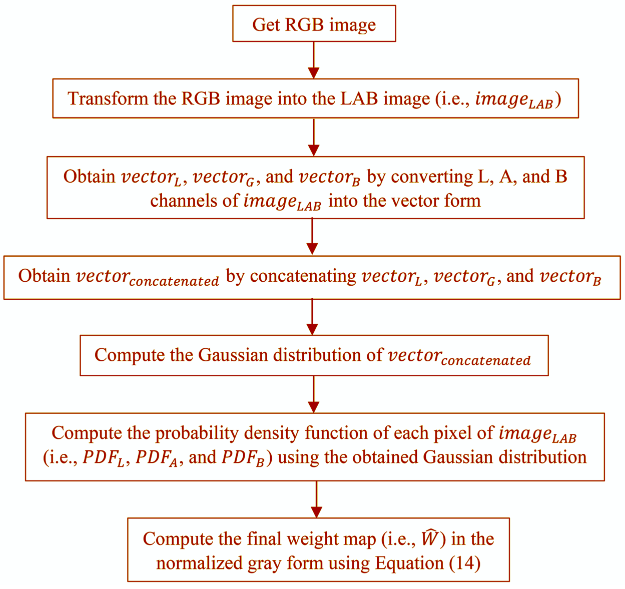
Get RGB image
Transform the RGB image into the LAB image (i.e., $image_{LAB}$)
Obtain $vector_L$, $vector_G$, and $vector_B$ by converting L, A, and B channels of $image_{LAB}$ into the vector form
Obtain $vector_{concatenated}$ by concatenating $vector_L$, $vector_G$, and $vector_B$
Compute the Gaussian distribution of $vector_{concatenated}$
Compute the probability density function of each pixel of $image_{LAB}$ (i.e., PDF_L, PDF_A, and PDF_B) using the obtained Gaussian distribution
Compute the final weight map (i.e., $\hat{W}$) in the normalized gray form using Equation (14)

Figure 2. The main steps used for obtaining the proposed weight map ($\hat{W}$).

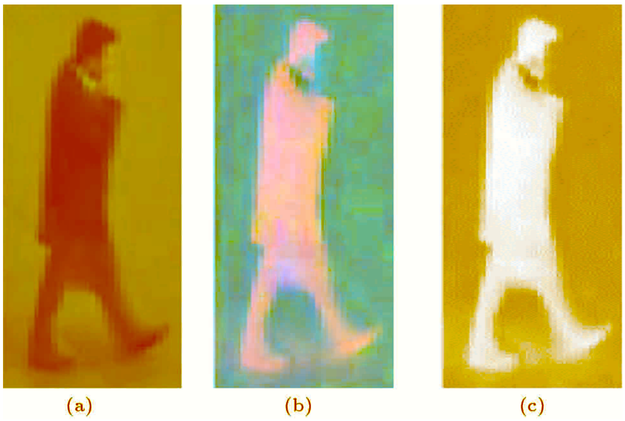

Figure 3. (a) A sample image, (b) the pixels probabilities (i.e., image W_{LAB}), and, (c) the normalized gray image (i.e., $\hat{W}$) obtained from W_{LAB}.

As can be seen in Figure 3, higher values (i.e., the brighter pixels) mainly refer to the person's body. It shows that in LAB color space, the background pixels have fewer probability values in the obtained Gaussian distribution compared to pixels associated with the person's body. This distinction is more obvious from the normalized gray image (i.e., $\hat{W}$) shown in Figure 3(c).

We use the GOG descriptor in a local form by considering $\hat{W}$ and the Pixels Saliency Map (PSM) [36] as follows:

(a) Weighing the raw extracted features introduced in Eq. (5), i.e., F_i, using $\hat{W}$ via Eq. (15):

$$\widehat{F}_i = \hat{W}_i \times F_i, \tag{15}$$

where $\hat{W}_i$ is the weight of the pixel i. According to this equation, $\hat{W}_i$ is used for weighing the feature vector of pixel i (i.e., F_i).

(b) Weighing the obtained Gaussian patches, which are utilized for obtaining Gaussian regions using the PSM. For this goal, the w_s in Eqs. (9) and (10) are substituted with the PSM.

A sample image from 3DPeS [53] and the corresponding PSM, and $\hat{W}$ are shown in Figure 4. As can be seen in this figure, the map using PSM and $\hat{W}$ mostly contains higher values for the pixels forming the body parts in comparison to the background pixels.

We name the improved versions of GOG and HGD proposed in this research as IGOG and IHGD, respectively. Similar to GOG, the final feature vector obtained from IGOG is as follows:

$$\hat{Z}_{RGB} = \left[\hat{z}_{\mathcal{G}1}^T, \hat{z}_{\mathcal{G}2}^T, \hat{z}_{\mathcal{G}3}^T, \hat{z}_{\mathcal{G}4}^T, \hat{z}_{\mathcal{G}5}^T, \hat{z}_{\mathcal{G}6}^T, \hat{z}_{\mathcal{G}7}^T\right]^T, \tag{16}$$

$$\hat{Z}_{Fusion} = \left[\hat{Z}_{RGB}^T, \hat{Z}_{LAB}^T, \hat{Z}_{HSV}^T, \hat{Z}_{nRGB}^T\right]^T, \tag{17}$$

where $\hat{z}_{\mathcal{G}i}^T$ represents the feature vector for region $\mathcal{G}i$. $\hat{Z}_{RGB}^T$ is the extracted feature vector using the raw features in Eq. (5). Also, $\hat{Z}_{LAB}^T$, $\hat{Z}_{HSV}^T$, and $\hat{Z}_{nRGB}^T$ denote the extracted feature vectors by substituting the RGB information (i.e., x_R, x_G, and x_B) used in Eq. (5) with the components of LAB, HSV, and nRGB. A similar approach is considered to obtain the final feature vector from IHGD. Note that, in our proposed weighing map, the pixels corresponding to objects or salient regions have a higher probability compared to other pixels. Indeed, the values of pixels belonging to objects or salient regions are usually different from the pixel values associated with the background and body. It leads to obtaining different PDFs for pixels corresponding to objects or salient regions from the other pixels. Hence, extracting raw features from images using our proposed weighing map leads to a higher focus on the pixels of carrying objects and salient regions compared to the other pixels.

In Figure 5, four sample images and their corresponding weight map (i.e., $\hat{W}$) obtained from our proposed weighing mechanism are shown. Also, sample images from the same person captured by another camera and their corresponding weight map are shown in Figure 6. Each column of these figures is related to a sample image.

As can be seen in these figures, the pose and background of the same person change in the field of view of different cameras. However, our proposed weighing mechanism approximately associates similar weights to the various parts of the same person in the two different views.

Similar to [1,2], the extracted feature vectors using our proposed weighing method in Eq. (17) are then used in an XQDA distance metric learning to simultaneously learn a low-dimensional discriminative subspace and a distance metric. After training the XQDA, the gallery images will be ranked according to their distance from the probe images.

4. Experimental results

In this section, the performance of the proposed re-identification approach is evaluated using Rank-k ($k = 1, 5, 10, 20$) measurement. Rank-k is a measure that can be used to evaluate and compare the performance of various re-identification methods. In this measure, k

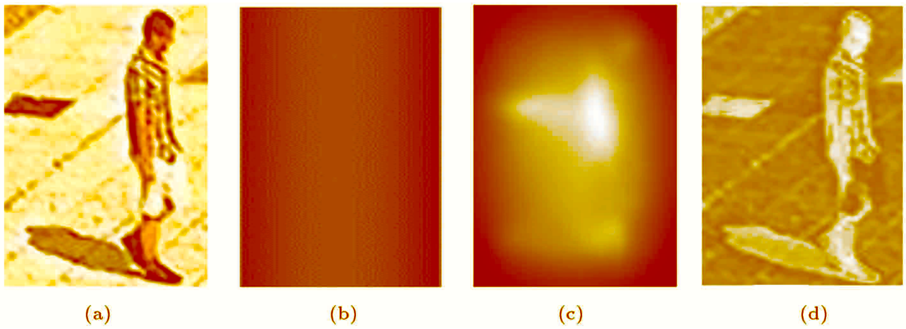

Figure 4. (a) A sample image and (b) the weight map using (c) PSM, and (d) $\hat{W}$ [32].

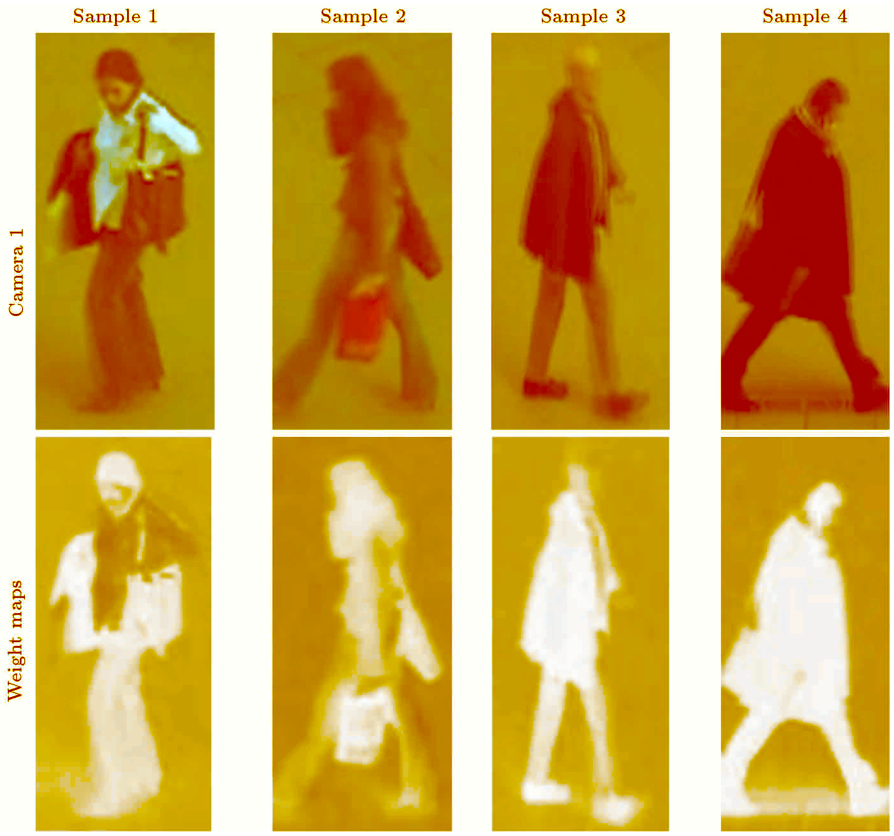

Figure 5. A set of sample images and their weight map obtained from our proposed approach.

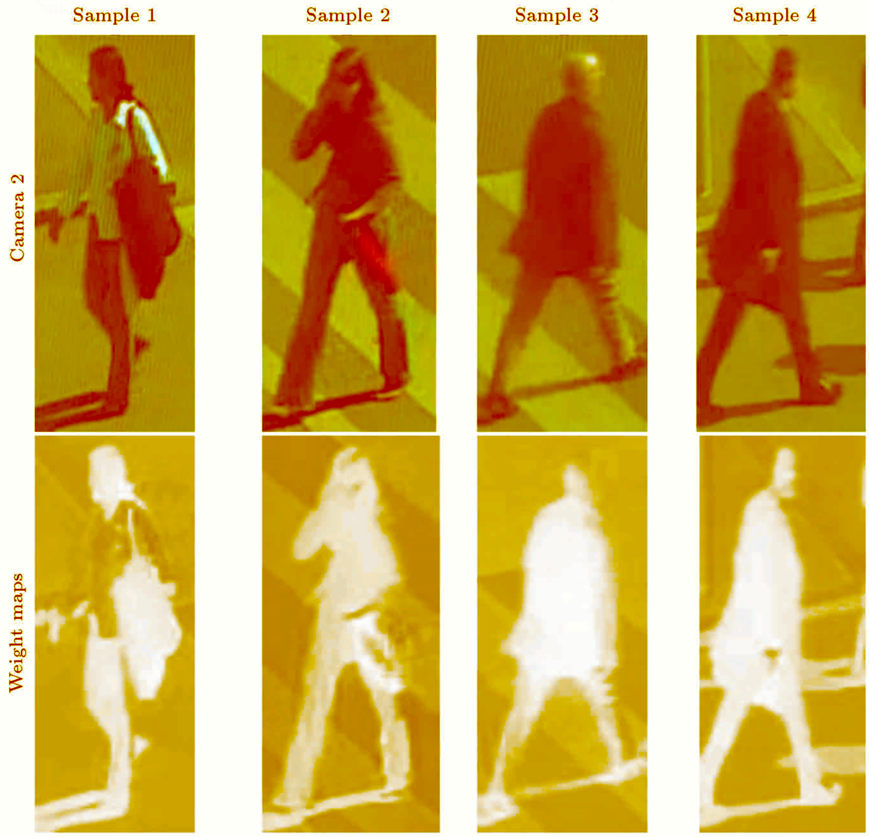

Figure 6. Another set of sample images and their weight map obtained from our proposed approach.

Table 1. Performance of our proposed approach and the methods proposed in [1,2].

		CUHK03 labeled	**VIPER**	**CUHK01 ($m = 1$)**	**CUHK01 ($m = 2$)**	**PRID 450 s**	**GRID**
Rank 1	IGOG	70.1	50.6	58.7	68.8	70.3	**26.7**
	classic GOG, (2016)	67.3	49.7	57.9	67.3	68.4	24.7
	IHGD	**71.2**	**51.9**	**61.4**	**72.7**	**72.8**	26.6
	classic HGD, (2019)	68.9	50.0	59.0	70.3	70.4	25.6
Rank 5	IGOG	91.8	79.77	79.9	88.0	89.4	**48.6**
	classic GOG, (2016)	91.0	9.7	79.2	86.9	88.8	47.0
	IHGD	92.1	**80.5**	**82.3**	**90.1**	**91.5**	48.0
	classic HGD, (2019)	**92.2**	79.5	79.7	87.9	91.2	46.7
Rank 10	IGOG	**96.3**	88.3	87.1	92.8	93.8	58.7
	classic GOG, (2016)	96.0	88.7	86.2	91.8	94.5	58.4
	IHGD	96.2	**89.1**	**88.7**	**94.0**	**95.0**	**59.4**
	classic HGD, (2019)	96.0	88.9	86.2	92.2	94.8	58.4
Rank 20	IGOG	**98.7**	95.2	92.6	96.6	96.8	68.2
	classic GOG, (2016)	–	94.5	92.1	95.9	**97.8**	**69.0**
	IHGD	98.6	**95.4**	**93.6**	**96.9**	97.6	68.9
	classic HGD, (2019)	98.7	94.6	92.0	95.8	97.6	68.5

determines the number of top matches with the correct answer [54]. Hence, for $k = 1$, Rank-k is the strictest measure, whereas, for $k > 1$, this measure permits some error [55].

The CUHK03 [56], VIPER [57], CUHK01 [58], PRID450s [59], and GRID [60] datasets were used to compare the performance of the proposed approach with previously reported results in [1,2].

The GRID database contains 1275 images of 1025 individuals captured by eight cameras applied in a busy underground station. The VIPeR and PRID450s databases, respectively, contain 1,264 images of 632 individuals and 900 images of 450 individuals. The images of the VIPeR and PRID450s databases were captured in two different camera views, where they involved one image of each person in each camera view. Hence, in GRID, VIPeR, and PRID450s, the performance of our proposed approach is evaluated using single-shot matching. The CUHK03 database involves 13,164 images of 1,360 individuals captured by ten cameras, where an average of 4.8 images of each person were captured in each camera view. In this paper, the manually cropped (labeled) images of this database are used. Besides, the performance of the comparing approaches is evaluated using multi-shot matching. The CUHK01 database contains 3,884 images of 971 individuals captured by two cameras. In this database, two images of each person were captured in each camera view. Hence, in Table 1, the performance of our proposed approach on the CUHK01 database is reported with single-shot matching ($M = 1$) and multi-shot ($M = 2$) matching.

Table 1 depicts the performance of these re-identification methods. In this table, the bolded results are the most accurate ones in each ranking order from the corresponding dataset. In addition, each ranking in Table 1 denotes the obtained results using our proposed weighing approach applied on the GOG descriptor (i.e., IGOG), the GOG method, the obtained results using our proposed weighing approach applied on HGD descriptor (i.e., IHGD), and the HGD method, respectively.

As shown in Table 1, the proposed approach achieves more accurate results compared to classic GOG and classic HGD for ranks 1, 10, and 20 on the CUHK03 dataset; for all ranks on VIPER and CUHK01 datasets; for ranks 1, 5, and 10 on PRID450s and GRID datasets.

Also, in Tables 2 to 6, we compare the performance of our proposed approach with other state-of-the-art methods on the CUHK03, VIPER, CUHK01, PRID450s, and GRID datasets, respectively. Note that most of the re-identification approaches reported their accuracy on the CUHK03 dataset only in rank 1; hence, in Table 2, we compare the performance of our people re-identification approach with the other

Table 2. Comparing the performance of our proposed approach with the state-of-the-art methods on CUHK03.

Approaches	Rank 1 (%) on CUHK03 (labeled)
Sun et al. [22] (2017)	40.9
Zheng et al. [23] (2018)	36.9
Li et al. [38] (2018)	44.4
Yu et al. [24] (2020)	53.9
Wu and Gao [27] (2020)	69.6
IGOG	70.1
IHGD	**71.2**

methods in rank 1. As shown in these tables, our proposed approach outperforms the comparing methods on CUHK03 (rank 1) and all ranks on VIPER and CUHK01 datasets, as well as in ranks 1, 5, and 10 on PRID450s and GRID datasets.

In general, considering the comparisons between the obtained results from our proposed re-identification approach and the results reported in the state-of-the-art methods on various datasets, our proposed re-identification approach improves the accuracy of person re-identification as it tunes the effect of each pixel in the extracted features proportional to pixels association with background or person's body.

Considering a sample image in the size of 128 × 48 as well as using a MATLAB implementation executed on a 2.6 GHz Intel Core i7 CPU, the computational times of our proposed weight map, PSM, and classic GOG were respectively obtained 0.0161, 0.2675, and 0.2673 seconds. Accordingly, the execution time of the IGOG was obtained at 0.5509 seconds. Considering these execution times, the overhead of our proposed weight map on the classic GOG descriptor can be neglected. Meanwhile, in most of the existing datasets used for people re-identification, the person is located in the middle of the images. Hence, PSM can be neglected for most of the images. Note that the classic HGD is similar to the classic GOG; hence, these points are true for IHGD. Consequently, the implementation time of the proposed method is reasonable.

The re-identification approaches introduced in [3,30] are simpler than our proposed approach, i.e., IGOG and IHGD in terms of computational and time complexity, whereas the accuracy of IGOG and IHGD is considerably higher than these approaches on the comparing datasets. Vishwakarma and Upadhyay [9] used the same feature extraction manner as GOG and HGD. Prates and Schwartz [10] proposed a Kernel X-CRC approach for dealing with the issue of appearance changes caused by different camera conditions. This approach was based on the appearance features extracted from GOG. Besides, the VS-SSL approach proposed in [19] is based on the GOG descriptor. Accordingly, Refs. [10,19] are more complex than GOG in terms of time complexity. In [36], color, shape, and texture characteristics were extracted both in the weighted form and non-weighted form. Then, a non-Euclidean metric was learned for each feature. Zhao et al. [20] considered people re-identification as a consistent iterative multi-view joint transfer learning optimal problem and solved the problem using the IALM algorithm. Zhou et al. [11] and Cao et al. [12] used a patch-wise graph-matching mechanism for training a set of patch-wise correspondence templates from positive image pairs with various pose-pair configurations. Also, in [32], considering some mid-level characteristics, a classifier

Table 3. Comparing the performance of our proposed approach with the state-of-the-art methods on VIPER.

Approaches	Ranks (%)			
	1	5	10	20
Layne et al. [32] (2012)	18.8	40.9	54.9	–
Zhao et al. [33] (2013)	26.7	50.7	62.4	76.4
Martinel et al. [36] (2014)	33.0	–	75.6	86.9
Liao et al. [3] (2015)	40.0	–	80.5	91.1
Vishwakarma and Upadhyay [9] (2018)	47.5	–	87.9	93.7
Leng [28] (2018)	32.9	60.3	73	–
Chu et al. [30] (2019)	49.0	74.1	84.4	93.1
Ren et al. [31] (2019)	42.1	64.0	73.4	–
Fang et al. [14] (2019)	43.8	79.2	87.2	94.9
Jia et al. [19] (2020)	44.8	72.3	79.3	86.1
IGOG	50.6	79.7	88.3	95.2
IHGD	**51.9**	**80.5**	**89.1**	**95.4**

Table 4. Comparing the performance of our proposed approach with the state-of-the-art methods on CUHK01 ($m = 2$).

	Ranks (%)			
Approaches	**1**	**5**	**10**	**20**
Liao et al. [3] (2015)	63.2	–	90.8	94.9
Vishwakarma and Upadhyay [9] (2018)	54.5	–	83.5	90.5
Fang et al. [14] (2019)	69.2	87.8	93.2	97.1
Prates and Schwartz [10] (2019)	63.1	82.7	89.0	94.6
Zhao et al. [20] (2020)	68.4	86.3	93.6	96.8
IGOG	68.8	88.0	92.8	96.6
IHGD	**72.7**	**90.1**	**94.0**	**96.9**

Table 5. Comparing the performance of our proposed approach with the state-of-the-art methods on Prid450s.

	Ranks (%)			
Approaches	**1**	**5**	**10**	**20**
Liao et al. [3] (2015)	62.6	85.6	92.0	96.6
Vishwakarma and Upadhyay [9] (2018)	62.4	-	93.5	96.9
Zhou et al. [11] (2018)	58:4	77.6	4.3	89.8
Leng [28] (2018)	38.2	67.1	76.0	–
Ren et al. [31] (2019)	60.6	82.8	90.8	–
Zhou et al. [13] (2019)	70:9	89.1	93.5	96.5
Jia et al. [19] (2020)	68.2	90.2	94.9	**98.0**
IGOG	70.3	89.4	93.8	96.8
IHGD	**72.8**	**91.5**	**95.0**	97.6

Table 6. Comparing the performance of our proposed approach with the state-of-the-art methods on GRID.

	Ranks (%)			
Approaches	**1**	**5**	**10**	**20**
Liao et al. [3] (2015)	16.6	–	41.8	52.4
Vishwakarma and Upadhyay [9] (2018)	23.7	–	58.2	68.1
Fang et al. [14] (2019)	20.4	**59.4**	**69.9**	**82.0**
IGOG	**26.7**	48.6	58.7	68.2
IHGD	26.6	48.0	59.4	68.9

was trained for each characteristic using a dataset. Using the mentioned training steps in these approaches brings more time complexity compared to the IGOG and IHGD, which don't use training steps. In [24], a number of investigators were simulated for each probe image by combining some appearance-based feature extraction approaches with various metric learning methods in pairs. Besides, the saliency learning and matching framework proposed in [33] computed a saliency map for each image in the gallery and probe sets by comparing the patches of images with a specified reference set. In [24] and [33], respectively, simulating a number of investigators for each probe images, and comparing each patch of the image with the patches that existed in the images of the reference set to make these approaches more complex than GOG and HGD in terms of computational and time complexity. Meanwhile, some re-identification approaches are based on training deep neural networks. Training these kinds of networks brings computational and time complexity. The re-identification approaches proposed in [14,28,31] used some appearance features such as color naming and features extracted from a deep network. The re-identification approach introduced in [14] is better than our IGOG and IHGD in ranks 5, 10, and 20 on the GRID dataset, whereas the IGOG and IHGD outperform in rank 1 on this dataset. Sun et al. [22] optimized the deep representation learning process using the SVD in order to extract global appearance features from CNN. The PAN was proposed in [23] based on the CNN feature maps, where it aligned the pedestrians within bounding boxes and learned the pedestrian descriptors simultaneously. Wu and Gao [27] fused the global and local features of a person's images through deep learning networks via their proposed MFF approach to extract discriminative appearance characteristics.

5. Conclusion

Accuracy in people re-identification directly affects the performance of surveillance systems. Many of the existing re-identification systems assume that the subject's appearance doesn't change during the tracking task. Re-identification systems may lose the subject during the tracking process due to real-time changes in the subject's appearance. In the proposed re-identification approach, descriptors that are robust to appearance changes, namely Gaussian Of Gaussian (GOG) and the Hierarchical Gaussian Descriptor (HGD), have been improved using our proposed pixel weighing scheme. The experimental results show that the proposed re-identification approach has a better performance compared to the other existing methods.

References

1. Matsukawa, T., Okabe, T., Suzuki, E., et al. "Hierarchical gaussian descriptor for person re-identification", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1363–1372 (2016).
2. Matsukawa, T., Okabe, T., Suzuki, E., et al. "Hierarchical gaussian descriptors with application to person re-identification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14 (2019).
3. Liao, S., Hu, Y., Zhu, X., et al. "Person re-identification by local maximal occurrence representation and metric learning", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206 (2015).
4. Liao, S., Zhao, G., Kellokumpu, V., et al. "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes", In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1301–1306 (2010).
5. Ullah, M., Ahmed, M., Cheikh, F.A., et al. "A hierarchical feature model for multi-target tracking", In *IEEE International Conference on Image Processing (ICIP)*, pp. 2612–2616 (2017).
6. Mallat, S.G. and Zhang, Z. "Matching pursuits with time-frequency dictionaries", *IEEE Transactions on Signal Processing*, **41**(12), pp. 3397–3415 (1993).
7. Kong, J., Teng, Z., Jiang, M., et al. "Video-based person re-identification with parallel spatial-temporal attention module", *Journal of Electronic Imaging*, **29**(1), pp. 1–17 (2020).
8. Wu, A., Zheng, W.S., and Lai, J.H. "Robust depth-based person re-identification", *IEEE Transactions on Image Processing*, **26**(6), pp. 2588–2603 (2017).
9. Vishwakarma, D.K. and Upadhyay, S. "A deep structure of person re-identification using multi-level gaussian models", *IEEE Transactions on Multi-Scale Computing Systems*, **4**(4), pp. 513–521 (2018).
10. Prates, R. and Schwartz, W.R. "Kernel cross-view collaborative representation based classification for person re-identification", *Journal of Visual Communication and Image Representation*, **58**, pp. 304–315 (2019).
11. Zhou, Q., Fan, H., Zheng, S., et al. "Graph correspondence transfer for person re-identification", In *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**(1), pp. 7599–7606 (2018).
12. Cao, Z., Simon, T., Wei, S., et al. "Realtime multi-person 2D pose estimation using part affinity fields", In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
13. Zhou, Q., Fan, H., Yang, H., et al. "Robust and efficient graph correspondence transfer for person re-identification", *IEEE Transactions on Image Processing*, **30**, pp. 1623–1638 (2019).
14. Fang, J., Zhang, R.F., and Jiang, F. "Sample specific multi-kernel metric learning for person re-identification", In *2nd IEEE International Conference on Electrical and Electronic Engineering*, Atlantis Press (2019).
15. Zhao, R., Ouyang, W., and Wang, X. "Unsupervised salience learning for person re-identification", In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3586–3593 (2013).
16. An, L., Kafai, M., Yang, S., et al. "Reference-based person re-identification", In *AVSS*, pp. 244–249 (2013).
17. Yang, Y., Yang, J., Yan, J., et al. "Salient color names for person re-identification", In *13th European Conference on Computer Vision (ECCV)*, pp. 536–551 (2014).
18. Ahmed, E., Jones, M., and Marks, T.K. "An improved deep learning architecture for person re-identification", In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015).
19. Jia, J., Ruan, Q., Jin, Y., et al. "View-specific subspace learning and re-ranking for semi-supervised person re-identification", *Pattern Recognition*, **108**, 107568 (2020).
20. Zhao, C., Wang, X., Zuo, W., et al. "Similarity learning with joint transfer constraints for person re-identification", *Pattern Recognition*, **97**, 107014 (2020).
21. Xu, Y., Fang, X., Wu, J., et al. "Discriminative transfer subspace learning via low-rank and sparse representation", *IEEE Trans. Image Process*, **25**(2), pp. 850–863 (2016).

22. Sun, Y., Zheng, L., Deng, W., et al. "Svdnet for pedestrian retrieval", In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3800–3808 (2017).

23. Zheng, Z., Zheng, L., and Yang Y. "Person alignment network for large-scale person re-identification", *IEEE Transactions on Circuits and Systems for Video Technology*, **29**(10), pp. 3037–3045 (2018).

24. Yu, Y., Liang, C., Ruan, W., et al. "Crowdsourcing-based ranking aggregation for person re-identification", In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1933–1937 (2020).

25. Sun, Y., Zheng, L., Yang, Y., et al. "Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)", In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 480–496 (2018).

26. He, K., Zhang, X., Ren, S., et al. "Deep residual learning for image recognition", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016).

27. Wu, S. and Gao, L. "Multi-level joint feature learning for person re-identification", *Algorithms*, **13**(5), pp. 111–129 (2020).

28. Leng, Q. "Co-metric learning for person re-identification", *Advances in Multimedia* (2018).

29. Bromley, J., Bentz, J.W., Bottou, L., et al. "Signature verification using a "siamese" time delay neural network", *International Journal of Pattern Recognition and Artificial Intelligence*, **7**(4), pp. 669–688 (1993).

30. Chu, H., Qi, M., Liu, H., et al. "Local region partition for person re-identification", *Multimedia Tools and Applications*, **78**, pp. 27067–27083 (2019).

31. Ren, Q.Q., Tian, W.D., and Zhao, Z.Q. "Person re-identification based on feature fusion", In *International Conference on Intelligent Computing*, pp. 65–73 (2019).

32. Layne, R., Hospedales, T.M., and Gong, S. "Towards person identification and re-identification with attributes", In *European Conference on Computer Vision*, pp. 402–412 (2012).

33. Zhao, R., Ouyang, W., and Wang, X. "Unsupervised salience learning for person re-identification", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3586–3593 (2013).

34. Zhao, R., Oyang, W., and Wang, X. "Person re-identification by saliency learning", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **39**(2), pp. 356–370 (2016).

35. Heller, K., Svore, K., Keromytis, A.D., et al. "One class support vector machines for detecting anomalous windows registry accesses", ICDM Workshop on Data Mining for Computer Security (2003).

36. Martinel, N., Micheloni, C., and Foresti, G.L. "Kernelized saliency-based person re-identification through multiple metric learning", *IEEE Transactions on Image Processing*, **24**(12), pp. 5645–5658 (2015).

37. Harel, J., Koch, C., and Perona, P. "Graph-based visual saliency", In *Advances in Neural Information Processing Systems*, pp. 545–552 (2007).

38. Li, W., Zhu, X., and Gong, S. "Harmonious attention network for person re-identification", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2285–2294 (2018).

39. Li, T., Sun, L., Han, C., et al. "Salient region-based least-squares log-density gradient clustering for image-to-video person re-identification", *IEEE Access*, **6**, pp. 8638–8648 (2018).

40. Susan, S. and Kumar, A. "Auto-segmentation using mean-shift and entropy analysis", In 2016 3rd *International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 292–296 (2016).

41. Ashizawa, M., Sasaki, H., Sakai, T., et al. "Least-squares log-density gradient clustering for riemannian manifolds", In *Artificial Intelligence and Statistics*, pp. 537–546 (2017).

42. Cheng, M.M., Mitra, N.J., Huang, X., et al. "Global contrast based salient region detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **37**(3), pp. 569–582 (2015).

43. Amari, S. and Nagaoka, H. "Methods of information geometry", *Translations of Mathematical Monographs*, American Mathematical Society, **191**, pp. 269–342 (2007).

44. Li, P., Wang, Q., and Zhang, L. "A novel earth mover's distance methodology for image matching with gaussian mixture models", In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1689–1696 (2013).

45. Arsigny, V., Fillard, P., Pennec, X., et al. "Geometric means in a novel vector space structure on symmetric positive-definite matrices", *SIAM Journal on Matrix Analysis and Applications*, **29**(1), pp. 328–347 (2007).

46. Lovric, M., Min-Oo, M., and Ruh, E.A. "Multivariate normal distributions parametrized as a riemannian symmetric space", *Journal of Multivariate Analysis*, **74**(1), pp. 36–48 (2000).

47. Moghaddam, B., Jebara, T., and Pentland, A. "Bayesian face recognition",*Pattern Recognition*, **33**(11), pp. 1771–1782 (2000).

48. Koestinger, M., Hirzer, M., Wohlhart, P., et al. "Large scale metric learning from equivalence constraints", In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295 (2012).

49. Alipanahi, B., Biggs, M., and Ghodsi, A. "Distance metric learning vs. fisher discriminant analysis", In *Proceedings of the 23rd National Conference on Artificial Intelligence*, **2**, pp. 598–603 (2008).

50. Septiarini, A., Hamdani, H., Hatta, H.R., et al. "Automatic image segmentation of oil palm fruits by applying the contour-based approach", *Scientia Horticulturae*, **261**, 108939 (2020).

51. Luong, Q.T. "Color in computer vision", In *Handbook of Pattern Recognition and Computer Vision*, World Scientific Publishing Co., Inc., pp. 311–368 (1993).

52. Le, K.N. "A mathematical approach to edge detection in hyperbolic-distributed and gaussian-distributed pixel-intensity images using hyperbolic and gaussian masks", *Digital Signal Processing*, **21**(1), pp. 162–181 (2011).

53. Baltieri, D., Vezzani, R., and Cucchiara, R. "3DPeS: 3D people dataset for surveillance and forensics". In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, pp. 59–64 (2011).

54. Swearingen, T. and Ross, A. "Lookalike disambiguation: improving face identification performance at top ranks". In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10508–10515 (2021).

55. Mortezaie, Z. and Hassanpour, H. "A survey on age invariant face recognition methods", *Jordanian Journal of Computers and Information Technology*, **5**(2), pp. 87–96 (2019).

56. Li, W., Zhao, R., Xiao, T., et al. "Deepreid: deep filter pairing neural network for person re-identification", In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 152–159 (2014).

57. Gray, D. and Tao, H. "Viewpoint invariant pedestrian recognition with an ensemble of localized features", In *European Conference on Computer Vision*, Berlin, Heidelberg, pp. 262–275 (2008).

58. Li, W., Zhao, R., and Wang, X. "Human re-identification with transferred metric learning", In *Asian Conference on Computer Vision*, pp. 31–44 (2012).

59. Roth, P.M., Hirzer, M., K¨stinger, M., et al. "Mahalanobis distance learning for person re-identification", In *Person Re-Identification*, pp. 247–267 (2014).

60. Loy, C.C., Xiang, T., and Gong, S. "Time-delayed correlation analysis for multi-camera activity understanding", *International Journal of Computer Vision*, **90**(1), pp. 106–129 (2010).

Biographies

Zahra Mortezaie received her BSc degree in Computer Engineering from the Shahrood University of Technology, Shahrood, Iran, in 2013 and an MSc degree in Computer Engineering from the Shahrood University of Technology, Shahrood, Iran, in 2017. She is currently a PhD student at Shahrood University of Technology, Shahrood, Iran. Her research interests include data mining, signal, image, and video processing, and artificial neural networks.

Hamid Hassanpour received his PhD from the Queensland University of Technology, Australia, in 2004. He is currently a Full Professor at the faculty of Computer Engineering, Sharood University of Technology, Shahrood Iran. His research interests include Image Processing, Signal Processing, and Data Mining. He has published over 220 journal and conference papers. He is the Editor-in-Chief of the Journal of Artificial Intelligence and Data Mining.

Azeddine Beghdadi received his PhD from the University of Pierre et Marie Curie (Paris 6). He is currently a Full Professor at the University of Paris 13 (Institut Galilée) Sorbonne Paris Cite. He published more than 280 international refereed scientific papers. His research interests include image quality enhancement/assessment, image/video compression, multimedia security, bio-inspired models for image analysis and processing, and physics-based image analysis and processing. Dr. Beghdadi is the founder of the European Workshop on Visual Information Processing (EUVIP). He is an Associate Editor of "Signal Processing: Image Communication" Journal, Elsevier, European Journal on Image and Video Processing, Springer Verlag, Journal of Electronic Imaging, SPIE Digital Library, and Mathematical Problems in Engineering Journal, Hindawi.