Person Re-Identification from Different Views based on Dynamic Linear Combination of Distances

Amani Elaoud, Walid Barhoumi, Hassen Drira, Ezzeddine Zagrouba

Noname manuscript No.
(will be inserted by the editor)

Person Re-Identification from Different Views based on Dynamic Linear Combination of Distances

Amani Elaoud [1] · **Walid Barhoumi** [1, 2] ·
Hassen Drira [3] · **Ezzeddine Zagrouba** [1]

Received: date / Accepted: date

Abstract Person re-identification from videos taken by multiple cameras from different views is a very challenging problem that has attracted growing interest in last years. In fact, the same person from significant cross-view has different appearances from clothes change, illumination, and cluttered background. To deal with this issue, we use the skeleton information since it is not affected by appearance and pose variations. The skeleton as an input is projected on the Grassmann manifold in order to model the human motion as a trajectory. Then, we calculate the distance on the Grassmann manifold, in order to guarantee invariance against rotation, as well as local distances allowing to discriminate anthropometric for each person. The two distances are thereafter combined while defining dynamically the optimal combination weight for each person. Indeed, a machine learning process learns to predict the best weight for each person according to the rank metric of its re-identification results. Experimental results, using challenging 3D (IAS-Lab RGBD-ID and BIWI-Lab RGBD-ID) and 2D (Prid-2011 and i-LIDS-VID) benchmarks, show

A. Elaoud
E-mail: amani.elaoud@fst.utm.tn

W.Barhoumi
E-mail: walid.barhoumi@enicarthage.rnu.tn

H.Drira
E-mail: hassen.drira@imt-lille-douai.fr

E.Zagrouba
E-mail: ezzeddine.zagrouba@fsm.rnu.tn
[1] Université de Tunis El Manar, Institut Supérieur d'Informatique, Research Team on Intelligent Systems in Imaging and Artificial Vision (SIIVA), LR16ES06 Laboratoire de recherche en Informatique, Modélisation et Traitement de l'Information et de la Connaissance (LIMTIC), 2 Rue Bayrouni, 2080 Ariana, Tunisia.
[2] Université de Carthage, Ecole Nationale d'Ingénieurs de Carthage (ENICarthage), 45 Rue des Entrepreneurs, 2035 Tunis-Carthage, Tunisia.
[3] IMT Lille Douai, Univ. Lille, CNRS, UMR 9189 – CRIStAL – Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France.

that the proposed method can boost re-id ranking thanks to its ability to define the optimal weight for each person independently of view and pose changes.

Keywords Re-Identification · Manifolds · Weighted distances · Human Skeleton · Random Forest

1 Introduction

Re-identification (re-id) is the task of recognising a person from multiple cameras and it aims to identify a person of interest in different locations' tracks from different cameras. Recently, re-id has become a popular research field thanks to its wide applications in many areas, such as criminal spotting [36], event detection [16] and video surveillance [44]. However, different factors make person re-id a very challenging task, such as pose variation, viewpoint change, clothing changing, lighting and occlusion. Several works have been proposed in the literature in order to deal with the issue of person re-identification from images and videos. For image as an input, body shape feature representation can be effectively generated for clothing variations within the framework of retrieval [49] and person re-id [22]. For instance, in [50], a cross-domain attribute representation based on Convolutional Neural Networks (CNN) has been validated within the framework of person re-identification. In fact, the investigated images have been captured by six different cameras and each camera is treated as a domain and each domain is considered as a target domain in turn. Each image has been annotated by 108 attributes (gender, wearing long hair...) and a classifier has been trained over the images of some cameras. The trained classifier is thereafter used in order to re-identify persons' images captured from other cameras. Furthermore, Li et al. [21] have adopted deep filter pairing neural network for person re-identification, and Wang et al. [40] have investigated a joint learning framework for re-id while unifying Single-Image Representation (SIR) and Cross-Image Representation (CIR) via convolutional neural networks. More recently, a multi-granularity image-text alignment has been proposed for person re-identification [31].

Nevertheless, in spite of the success of modern deep learning architectures [12] within the framework of person re-identification from static images, video-based person re-id has attracted much attention. In fact, videos contain space-time as well as motion information [11] allowing richer information than independent images for an accurate re-identification. Thus, we are interested herein in the video-based recognition of persons from multiple cameras. This consists to identify a person of interest in different locations' tracks from different cameras. In fact, we present a model that indicates the identity of a testing person in different views. More precisely, we propose an accurate method based on measuring the similarity between trajectories for two persons in order to identify a person of interest across different camera views on different locations and at different times. It is worth noting that this paper is an extension of our previous conference paper [8] towards the general framework of re-identifying

a given person among a large set of persons over various camera views. The main contribution of the proposed method resides in adopting machine learning, and more precisely random forest, for the dynamic linear combination of distances (Grassmann distance and local distance) according to each person. Indeed, we combine different weights in order to define dynamically the optimal weight for each person while considering shape as well as anthropometric differences between persons. Since the same person observed in different camera views can have significant variations, we propose to extract the adequate weight for each person from one view and the random forest predicts the best weights for the testing persons from other views. The main goal is to learn how to automatically choose the appropriate weight in a dynamic manner through pairwise constraints between global and local distances and each person to be re-identified, without being obliged to broadly make the choice for a dataset of persons' videos. The realized tests on 3D datasets as well as on 2D datasets have proved the effectiveness of the proposed person re-identification method.

The rest of this paper is organized as follows. In section 2, we briefly review the related work on video-based person re-identification. In section 3, we describe the proposed method, and experimental results are discussed in section 4 in order to demonstrate the effectiveness of the proposed method. Then, a brief discussion is presented in section 5. Finally, in section 6, we conclude the proposed work and present some ideas for future studies.

2 Related Work

Relevant methods on video-based person re-identification can be regrouped according either to their inputs (RGB *vs.* RGB-D) or to their content description approaches (hand-crafted *vs.* deep-learned). In fact, first attempts have investigated RGB-based features such as color [46], shape [1] and texture [41]. For instance, Zhang et al. [53] have extracted rich spatial-temporal information for feature representation within the context of video-based re-id. Likewise, a graph representation learning approach has been introduced in [45] using pose alignment connection and feature affinity connection, which models the intrinsic relations between graph nodes. McLaughlin et al. [28] have investigated features that are extracted from consecutive video frames through a CNN model. Liu et al. [24] have learned both spatial features and motion context using accumulative motion context network. In [20], a global-local temporal representation has been exploited to multi-scale temporal cues in videos. Recently, Zhao et al. [54] have proposed a method for the similarity learning with joint transfer constraints. Generally, video-based re-identification faces many severe challenges, particularly viewpoint changing [7] and cluttered background [10]. Moreover, even under the same viewpoint, images belonging to a person may differ considerably due to the dramatic variations caused by variable illumination, pose and occlusion [18]. To deal with these issues, notably the fact that

the same person may show different appearances, many recent works have used RGB-D information, in order to profit from its insensitivity to pose and color variations. For instance, in [4], authors have studied person identification in indoor environments using skeletons detected by a Kinect V2 device along with wearable devices equipped with inertial sensors. Similarly, Takac et al. [37] have investigated people identification in a small home RGB-D camera network based on a combination of the SVM and the naive Bayes classifiers. Differently, Liang et al. [23] have exploited 3D locations of joints from appearance, instead of the pose obtained by Kinect, and corresponding confidence metrics of two videos in order to compute their confidence-weighted pose distance. Moreover, in [2], data acquired from RGB-D camera presenting a set of 3D soft-biometric cues have been used. Furthermore, a biometric metric learning, based on human skeleton by defined joint distances, has been proposed in [43]. Another work has investigated depth and skeleton information in order to extract the following features: histograms of local binary patterns, local derivative patterns and local tetra patterns [14]. Imani et al. [15] have exploited RGB, depth and skeleton information from RGB-D sensors in order to exploit complementarity of features extracted by different modalities [39]. Similarly, in [8], re-id has been based on the analysis of skeleton shape trajectories using different distances. However, given the lack of availability of 3D datasets, many studies have tried to extract skeletons from RGB videos. In fact, considering the lack of 3D datasets, various works have focused on 3D human pose estimation and predicting landmarks locations for a given sequence of RGB images. These studies proposed efficient tools for pose estimation from multiple images, captured by synchronized cameras, and even from a single image. This can be performed within the framework of single person pose estimation as well as of multi-person pose estimation. For instance, Yasin et al. [47] have proposed an elegant approach for 3D pose estimation from a single image. 3D pictorial structures have also been investigated for 3D pose estimation of multiple humans from multiple views [3]. For single person pose estimation, a deep architecture and a graphical model based on architecture with geometric relationships between body joint locations have been proposed in [38]. Differently, in [26], 3D joint locations from the 2D joints have been estimated by regression models. Recently, real-time multi-person 2D pose estimation has been proposed using part affinity fields [5]. In our case, we have adapted the algorithm of [5], which proved its effectiveness in obtaining the main landmarks from RGB videos within the framework of person re-id.

Furthermore, concerning the issue of content description for video-based re-identification, recent research works can be classified into two groups: hand-crafted features and deep-learned features. For instance, in order to improve the performance of video re-id, Liu et al. [25] have introduced a spatio-temporal appearance representation for video-based pedestrian re-identification, and the HOG3D descriptor with dense sampling has been adopted in [42] for identifying persons in video sequences. Furthermore, Paisitkriangkrai et al. [32] have used multiple low-level hand-crafted and high-level visual features [19]

for accurate person re-id. However, the second class of methods focuses on deep learning, such as [28], where Recurrent Neural Networks (RNN) have been used to overemphasize the temporal dependencies related to person's motion. Likewise, Chen et al. [6] have processed videos by using RNN in order to temporally aggregate spatial information extracted from CNN. Recently, RNN have been integrated with temporal spatial features for video-based person re-id [51]. In our case, we have opted for hand-crafted features based on distances comparison, since they are simpler, comprehensible and they have achieved good performance on small datasets. In fact, deep learning features outperform hand-crafted ones, but they require large annotated datasets for an accurate training and sometimes there is a problem of overfitting in the process of obtaining the features.

In this study, we propose an effective method based on hand-crafted features for person re-id in videos taken by multiple cameras from different views. In order to assure the highly efficient representation of human body structure and motion, we use the skeleton information since it is not affected by the appearance variation and the pose change. Moreover, given the lack of 3D datasets that are designed for person re-identification, we adapt an accurate algorithm to estimate 2D joints, simulating the skeletons, within RGB videos. This aims to guarantee that the proposed method can effectively deal with RGB datasets. In fact, we are particularly interested in person re-identification from 2D as well as 3D video sequences captured from two different camera views. The 3D human motion is presented as a weighted linear combination between distances on the Grassmann manifold, in order to guarantee invariance against rotation, and local distance between the joints of the skeleton in order to discriminate anthropometric for each person. Each person is dynamically characterized by the best combination of Grassmann and local distances according to its personal shape and anthropometric.

3 Proposed Method

In order to deal with the main issue of appearance changing, we propose herein a person re-id method based on the analysis of the skeleton motion. In fact, the skeleton information has proved to be more robust, than the RGB information, against the changing of the video acquirement environment (point of view, person clothes, illumination, background...). Fig. 1 illustrates some examples of RGB sequences, within the i-LIDS-VID dataset, for the same person with different appearances. Sequences in the same row are from the same person, nevertheless the appearance is very different given the changing of the video acquirement environment. The key idea of this work is to dynamically predict the best weight for combining the shape and the anthropometric measures of the skeleton along a sequence. Fig. 2 shows the outline of the proposed method. For the training phase, we have as an input the skeleton sequence of each studied person and the skeleton sequences of the rest of persons within a gallery of videos. For each person, we calculate its local distance ($Dist_{Local}$) as well as

Fig. 1 Examples of sequences from the i-LIDS-VID dataset. Sequences in the same row are from the same person (only a sample of six frames is shown for each sequence).

its global distance ($Dist_G$) with regard to all the other persons. Then, we test, for each person, all the linear combinations of its local and global distances with the remaining persons ($\alpha.Dist_{Local} + (1-\alpha).Dist_G$), $\alpha \in \{0, \ldots 1\}$. This results in a matrix $MatDist$ that synthesizes, for each person, all possible weighted combinations of its distances with the other persons in the studied dataset. The matrix represents the input for the machine learning process, based on random forest, which learns to predict the best weight for each person for one view according to the rank metric of its re-id results. During the testing phase, the local and global distances are computed for each testing person. Then, the trained machine learning model predicts the more adequate weight α^* for this person. This weight is thereafter adopted in order to rank the similarity of the testing person with regard to the persons forming the studied dataset. Moreover, in order to make the method applicable even for RGB data, we adapt the algorithm of [5] in order to predict landmarks simulating the skeleton of the input person. This algorithm uses 2D pose estimation from RGB data in order to deal with the problem of localizing anatomical keypoints allowing to detect body parts of individuals. It estimates landmarks' coordinates for each person in each image using CNN in order to jointly estimate confidence maps for body part detection. Examples of skeletons extracted from RGB images by the algorithm of [5] are shown in Fig. 3. Although the algorithm of [5] has some errors (missing or false part detection), like most of existing similar algorithms, it permits generally to estimate skeleton information that is robust to the variations of background, viewpoint and lighting.

3.1 Skeleton motion modeling

In order to retrieve the identity of a given probe sequence, we compare it to all sequences in the gallery. This comparison is performed on two different spaces. On the one hand, the shape of the skeleton at each frame of the se-

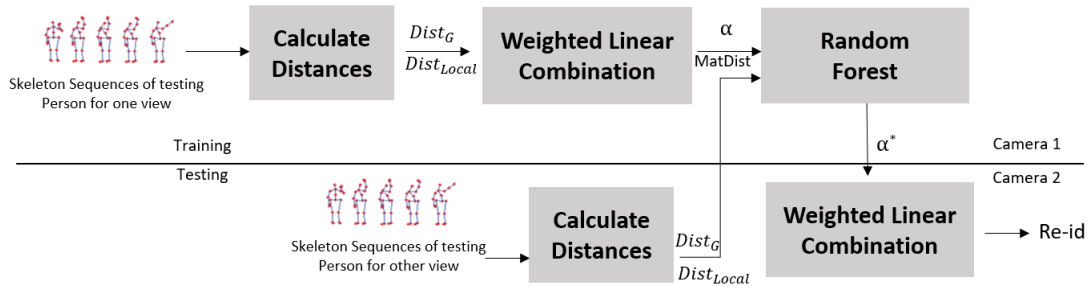

Fig. 2 Outline of the proposed re-identification method.

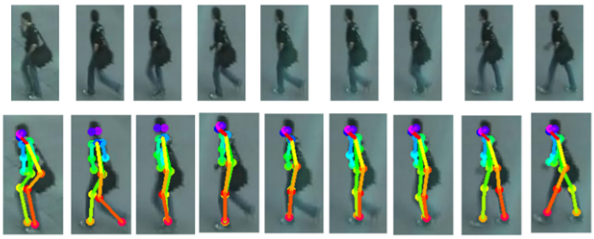

Fig. 3 2D landmarks estimation for body parts. The first row illustrates 2D images from an RGB video sequence and the corresponding skeletons are displayed in the second row.

quence is projected on the Grassmann manifold resulting to a trajectory on the non linear space. In fact, the projection is mainly performed using Singular Value Decomposition (SVD)-based orthogonalization [8]. Given a skeleton sequence S, we propose to model each frame as a point on the Grassmann manifold (a matrix) in order to transform the video on a trajectory T that links different points on the Grassmann manifold. Each point is represented by 3D coordinates of body joints or 2D coordinates (landmarks). Indeed, a manifold is a topological space that is locally similar to Euclidean space. Many existing works have used manifolds for re-identification such as the work of [27], which has presented each color that indicates the same feature vector as Gaussian distribution by focusing on the Symmetric Positive Definite (SPD) matrix manifold. Likewise, Fan et al. [9] have introduced person image mapped onto a hypersphere manifold for person re-identification. Recently, Zhao et al. [52] have proposed hyperspectral image unsupervised classification framework based on robust manifold matrix factorization and its out-of-sample extension. Generally, the skeleton data perform success with the projection on manifold space to facilitate the manipulation of the number of landmarks, what gives rise to the notion of manifolds performing dimensionality reduction. In our case, given its robustness against rotation, we have opted for the Grassmann manifold that is considered as a quotient space of the Stiefel manifold. The Stiefel manifold $V_{k,n}$ is the space whose points are k-frames in $\mathbb{R}^n$, where a set of k-orthonormal vectors in $\mathbb{R}^n$ is called a k-frame in $\mathbb{R}^n$ ($k \leq n$). Each point

on the Stiefel manifold $V_{k,n}$ can be represented as a $n \times k$ matrices X such that $X^T.X = I_k$, where I_k is the $k \times k$ identity matrix. Thus, the Grassmann manifold $G_{k,n-k}$ is the space whose points are k-planes or k-dimensional hyperplanes (containing the origin) in $\mathbb{R}^n$. An equivalent definition of the Grassmann manifold is as follows [48]. To each k-plane ν in $G_{k,n-k}$ corresponds a unique $n \times n$ orthogonal projection matrix P idempotent of rank k onto ν. If the columns of an $n \times k$ matrix Y spans ν, then, $YY^T = P$. Between two points U_1 and U_2 in $G_{k,n-k}$, there are $n-k$ principal angles in $\mathbb{R}^k$. The principal angles may be the inverse cosine of the singular values of ${U_1}^T U_2$. In fact, two points on the Grassmann manifold are equivalent only if one can be mapped into the other one by a rotation matrix. Thus, we calculate the distance $Dist_G$ on the Grassmann manifold, which represents the length of the shortest curve connecting two points on the Grassmann manifold (1). The trajectories' comparison on the Grassmann manifold includes a Dynamic Time Warping (DTW) step [35] in order to allow the invariance to the rate. The resulting distances represent the shape dissimilarities between the probe sequence T_1 and the gallery one T_2.

$$Dist_G(T_1,\ T_2) = \sum_{f=1}^{N} \theta_f^2, \tag{1}$$

where, N denotes the number of frames in the probe sequence T_1 and θ_f is the principal angle between two points (frames), $frame_1$ ($\in T_1$) and its DTW-based corresponding $frame_2$ ($\in T_2$), on the Grassmann manifold.

On the other hand, we perform a comparison of the anthropemetric measures, characterising each testing person, while measuring the lengths of skeleton parts. Fig. 4 shows an example of the twenty joints given by the skeleton and the lengths $d_1,\ d_2,\ \ldots,\ d_{19}$ of all segments connecting adjacent anatomical landmarks and illustrating the main joint connections of the human body driving the motion. Indeed, we compute distances between adjacent landmarks in order to obtain a vector of distances L ($= [d_1,\ d_2,\ \ldots,\ d_{19}]^T$) for each frame characterising each person. For each couple of skeletons ($\in T_1 \times T_2$), we compare two vectors L_1 and L_2 of 19 distances, and the local distance $Dist_{Local}$ (2) between the probe sequence T_1 and the gallery one T_2 is defined as the Frobenius norm $||\ ||_F$ of the differences between the M vectors of distances of the T_1 skeletons and the ones of their DTW-based corresponding skeletons in T_2. In fact, the comparison of sequences using this feature is performed using the dynamic time warping algorithm, similarly to the previous step, in order to ensure the invariance to the rate of execution. Finally, we combine the first distance on the Grassmann manifold $Dist_G$ and the second local distance $Dist_{Local}$ in order to find the best weighted linear combination (Algorithm 1).

$$Dist_{Local}(T_1,\ T_2) = ||Ske_1 - Ske_2||_F, \tag{2}$$

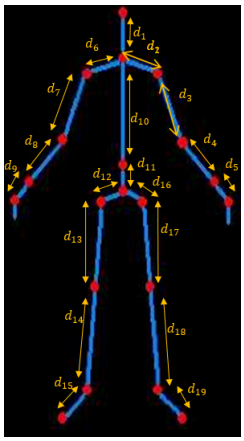

Fig. 4 Local features represented by the lengths of different parts of the skeleton.

where, Ske_1 is the $19 \times M$ matrix illustrating the 19-dimensional vectors of distances of the M skeletons within T_1, and Ske_2 denotes the vectors of distances of the T_2 skeletons that have been associated to the T_1 ones.

Algorithm 1 Weighted linear combination of distances

[1] **Input:** S_1 and S_2: Video sequences; α: Weight **Output:** $Dist$: Weighted linear combination of distances **Function: Weighted linear combination of distances**(S_1, S_2, α)
$frame_1 = 1$ to SizeOf(S_1) $T_1(frame_1) \leftarrow$ SVD $(S_1(frame_1))$ $frame_2$=1 to SizeOf(S_2) $T_2(frame_2) \leftarrow$ SVD $(S_2(frame_2))$
$D_1 \leftarrow Dist_G(DTW(T_1, T_2))$ // (Equation 1)
$frame_1 = 1$ to SizeOf(S_1) $L_1(frame_1) \leftarrow$ Compute distances between adjacent joints $(S_1(frame_1))$ $frame_2 = 1$ to SizeOf(S_2) $L_2(frame_2) \leftarrow$ Compute distances between adjacent joints $(S_2(frame_2))$
$D_2 \leftarrow Dist_{Local}(DTW(L_1, L_2))$ // (Equation 2)
$Dist \leftarrow \alpha.D_1 + (1 - \alpha).D_2$
return$(Dist\ )$ **end Function**

3.2 Dynamic linear combination of distances

The main contribution of this paper lies the adaptive linear combination, according to each person, of the Grassmann distance $Dist_G$ and the local distance $Dist_{Local}$ for more accurate person re-identification. The goal of this combination is to find a trade-off between these two distances considering the assessment of similarity between two trajectories according to the identity of the testing person (*i.e.* each person has his specific weight α^*). In fact, we have sampled the training sequences on two subsets $SeqTrain_1$ ($[E_1, \ldots, E_{n_1}]$) and $SeqTrain_2$ ($[Q_1, \ldots, Q_{n_2}]$), and we have tested different values $Val_\alpha(k)$ of the weight (*i.e.* 0.1, 0.2, ..., 0.9, 1) on the training set in order to find the

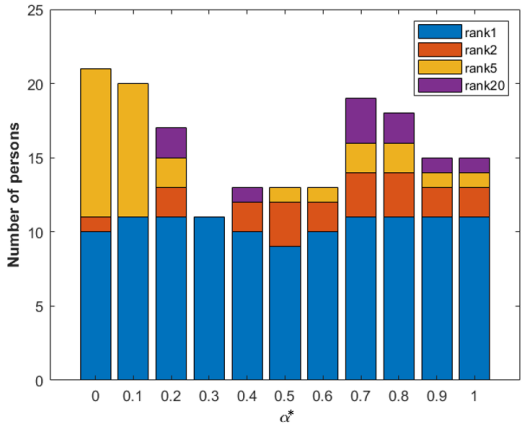

Fig. 5 Number of persons, within the 2D dataset iLIDS, for each value of α^*, for which it was ranked as first, second, fifth and twenty.

best weighted linear combination for each studied person. The selection of the tested values of α_i is strongly depending on the shape and anthropometric of the studied person. Thus, we have for each sequence E_i on camera1 the best weight α_i^*. In Fig. 5, we display for each α^* the number of persons, within the 2D dataset iLIDS, for which it was ranked as first, second, fifth and twenty. It is clear that no value of the weight α^* is the best for the entire set of the tested sequences, and even more, many least performing values of α^* on the entire set of tested sequences are top ranked for some individual sequences. Thus, there is not one weight that can be effectively applied for all persons.

Our insight has been inspired by this analysis that motivated us to investigate the possibility of selecting dynamically the appropriate α^* for each person according to his/her personal shape and motion, without being obliged to make this choice in a global manner for an entire dataset of sequences. Indeed, we propose an elegant model, based on machine learning, for predicting the weighted linear combination between the distance on the Grassmann manifold and the local distance. To this end, a random forest model is trained in order to learn how to extract automatically the best weight α^* for each treated person according to his shape as well as to his motion, which are described through his local and global distances with regard to others persons forming the studied dataset, for re-identifying persons within a precise set of persons' videos from different views. Then, we apply the identified values of α^* on the set of testing sequences in order to identify the testing sequences. These testing sequences $SeqTest$ are composed of videos performed by the same persons of the training set on camera2 $[V_1, \ldots, V_{m_1}]$. In fact, a dynamic linear combination of distances (Algorithm 2) is adopted to test different weights for each studied person and to use thereafter the random forest in order to obtain the predicted label of weight to identify the same person on camera2. We have used

different datasets which contain different camera views for each person. This can be very useful for various fields dealing with multi-view of cameras, such as the large variations between person videos captured by different cameras due to changes in illumination, pose, viewpoint and background.

Algorithm 2 Dynamic linear combination of distances

Input: $SeqTrain_1 = [E_1, \ldots, E_{n_1}]$: $Subset_1$ of training sequences (camera1); $SeqTrain_2 = [Q_1, \ldots, Q_{n_2}]$: $Subset_2$ of training sequences (camera1); $SeqTest = [V_1, \ldots, V_{m_1}]$: Set of testing sequences (camera2);
Val_α = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1] **Output:** Id($V_1, \ldots, V_{m1}$): Identity of testing persons ($V_1, \ldots, V_{m_1}$) **Function: Dynamic linear combination of distances** ($[E_1, \ldots, E_{n1}], [Q_1, \ldots, Q_{n2}], [V_1, \ldots, V_{m_1}], Val_\alpha$) i=1 to n_1 $RankMin \leftarrow$ 1000 k=1 to Size(Val_α) j=1 to n_2 $Dist_1(i,j) \leftarrow$ Weighted linear combination of distances($E_i, Q_j, Val_\alpha(k)$)
$Rank \leftarrow$ Recognition based on Nearest Neighbor $Dist_1(i,:)$ $Rank < RankMin$ $RankMin \leftarrow Rank$ $\alpha(i) \leftarrow Val_\alpha(k)$ $MatDis \leftarrow Dist_1(i,:)$
$\alpha^* \leftarrow$ Random Forest $(\alpha, MatDis)$ i=1 to m_1 j=1 to n_2 $Dist_2(i,j) \leftarrow$ Weighted linear combination of distances($V_i, Q_j, \alpha^*(i)$)
$Id_i \leftarrow$ Recognition based on Nearest Neighbor ($Dist_2(i,:)$) return(Id_i)
end Function

4 Results

We have tested the proposed re-id method on two challenging 3D datasets: IAS-Lab RGBD-ID [29] and BIWI-Lab RGBD-ID [29], as well as on two standard 2D datasets: iLIDS-VID [42] and PRID 2011 [13]. The IAS-Lab RGBD-ID dataset is an RGB-D dataset exclusively designed for re-identification from videos captured using RGB-D cameras. It includes 11 training sequences and 22 testing sequences of 11 different people using synchronized RGB images, depth images and skeletal data with three sets. The first one "Training" and the second one "TestingA" were recorded with people wearing different clothes. However, the third set "TestingB" was collected in a different room with the same clothes as in the first sequence. The BIWI RGBD-ID dataset is an RGB-D dataset also used for re-identification from RGB-D cameras. It is composed of 50 training sequences and 56 testing sequences of 50 different people, since 28 people out of 50 present in the training set have been recorded also in two testing videos each. This dataset is composed of a "Still" sequence and a "Walking" sequence collected in different days and in different locations while the subjects have been dressed differently. It is worth noting that we have used the full training set and the "Walking" testing set that contains dynamic skeleton data, similarly to the work of [34].

For the IAS-Lab RGBD-ID dataset, we have validated the proposed method using "Training" set and "TestingA" set, then "Training" set and "TestingB" set. We have started by computing for each studied person on "Training" set the best weight α^* according to his distances with the rest of persons. Then,

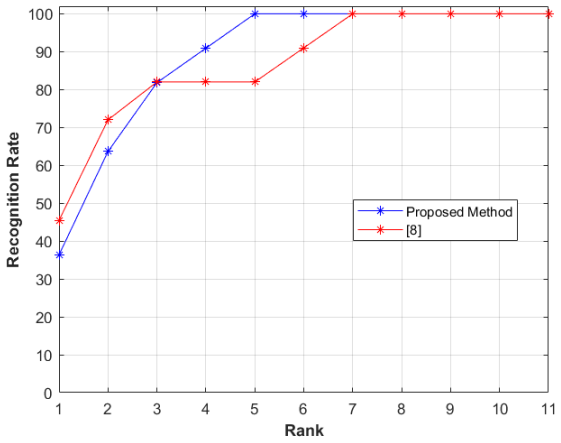

Fig. 6 Comparison of the proposed method against a relevant method from the state-of-the-art, on the IAS-Lab RGBD-ID (TestingA) dataset, in terms of Rank CMC (%).

we have applied the predicted weight α^* for the testing person in "TestingA" set in order to identify each person. Fig. 6 shows that the proposed method outperforms the work presented in [8] with value equals to 90.90% *vs.* 82% at rank 4 and it reaches a value of 100% at rank 5, whereas the work of [8] reaches the value of 100% only at rank 7. However, the work of [8] outperforms the proposed method on rank 1 with value equals to 45.45% *vs.* 36.36%. Thus, the proposed method records much higher matching rates than the work of [8] and for some other cases the work of [8] outperforms the proposed method. In Table 1, we conduct an extensive comparison with existing skeleton methods from the literature. It is clear that the proposed method outperforms in rank 1 the compared methods that are based on hand-crafted features ([30] and [33]) as well as the studied method based on deep learning of [55]. Furthermore, we have repeated the same action while using "Training" set and "TestingB" set (Fig. 7), similarly to the previous step. Fig. 7 shows that the suggested method and our previous work [8] are comparable in term of accuracy. In fact, the proposed method outperforms the work of [8] with value of 81.81% *vs.* 72.72% at rank 1, but the work of [8] outperforms the proposed method on the others ranks until we reach in the same rank (rank 9) the value 100%. We can conclude that the used data are on small datasets (11 persons) which do not promote machine learning to obtain good results. From table 1, we can deduce that the Proposed Method (PM) is more accurate in rank 1 to many recent skeleton-based methods ([30] and [33]) and it even outperforms the deep learning methods ([55] and [34]) on many datasets.

For the second used $3D$ dataset (BIWI-Lab RGBD-ID), we have used the "Training" set and the "Walking" set. We have identified for every person on "Training" set the best α^* to identify the studied person. Then, we have ap-

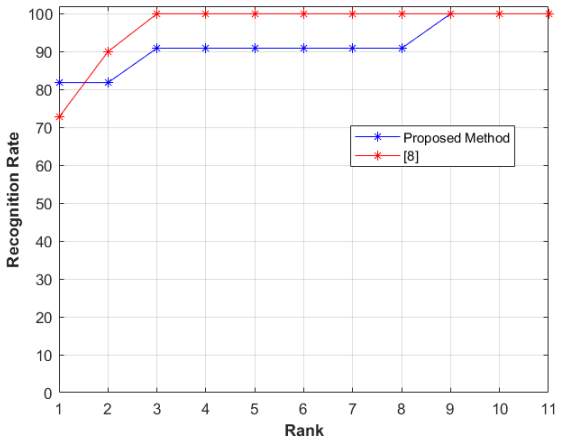

Fig. 7 Comparison of the proposed method against a relevant method from the state-of-the-art, on the IAS-Lab RGBD-ID (TestingB) dataset, in terms of Rank CMC (%).

Table 1 Comparison with existing skeleton-based methods, on the IAS-Lab RGBD-ID and the BIWI-Lab RGBD-ID datasets, in terms of Rank 1 (%).

Hand-crafted			
Methods	IAS-A	IAS-B	BIWI
Munaro et al. [30]	33.8	40.5	39.3
Elaoud et al. [8]	45.45	72.72	7.14
Pala et al. [33]	27.4	39.2	41.8
PM	36.36	81.81	39.3
Deep Learning			
Methods	IAS-A	IAS-B	BIWI
Zheng et al. [55]	34.4	30.9	36.1
Rao et al. [34]	56.1	58.2	59.1

plied the random forest in order to predict the best weight for the same studied person on the "Walking" set, according to the rank metric of re-identification results. Fig. 8 demonstrates that the proposed approach records much higher matching rates than the work of [8] when we use data with larger sizes. Furthermore, the results in Table 2 show that the proposed method is consistently better than the previous work [8], using the PRID-2011 dataset, with value of rank 1 equals to 87.02% *vs.* 83.14%, rank 5: 93.48% *vs.* 92.69%, rank 10: 95% *vs.* 95.50% and rank 20: 98.25% *vs.* 97.19%. In fact, there are sufficient training data available to predict the best weight. Using datasets with important sizes helps dynamic linear combination distances to perform accurate results. Fig. 9 shows that using the best weight for each person outperforms the fact of adopting the same weight for the entire studied dataset, given that each person has his distinctive motion and pace characteristics. It is worth noting that we have investigated the effect of several training parameters on machine learning random forest (Table 3).

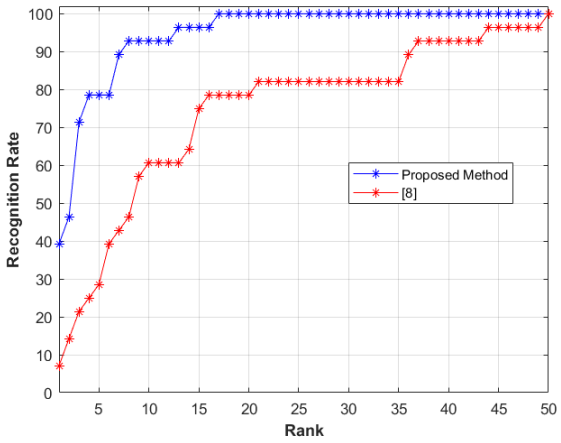

Fig. 8 Comparison of the proposed method against a relevant method from the state-of-the-art, on the BIWI-Lab RGBD-ID ("Walking") dataset, in terms of Rank CMC (%)

Table 2 Comparison of the proposed method with the work of [8], on the PRID-2011 dataset, in terms of Rank CMC (%).

CMC Rank	1	5	10	20
Elaoud et al. [8]	83.14	92.69	95.50	97.19
PM	87.02	93.48	95	98.25

Table 3 Training parameters.

Dataset	Number of training data	Number of trees	Depth
IAS-Lab RGBD-ID ("TestingA")	11×11	300	9
IAS-Lab RGBD-ID dataset ("TestingB")	11×11	100	9
BIWI-Lab RGBD-ID ("Walking")	28×50	1000	9
iLIDS-VID	150×150	1000	11
PRID 2011	200×200	1000	9

Furthermore, as $2D$ data, the PRID 2011 dataset is composed of 749 persons and is captured by two non-overlapping cameras, with sequences' lengths of 5 to 675 frames. Following the protocol used in [42], we have considered only the first 200 persons, who appear in both cameras. The second used $2D$ dataset is the iLIDS-VID dataset, which contains 300 persons, and each person is represented by two video sequences captured by non-overlapping cameras. For these experiments, each dataset was randomly split into 50% of persons for training and 50% of persons for testing. All experiments were repeated 10 times with different test/train splits and the results have been averaged in order to ensure stable results. The proposed method has been compared with the work of [42], which is based on a combination of HOG3D features and optic flow energy profile over each image sequence, as well as with a second relevant

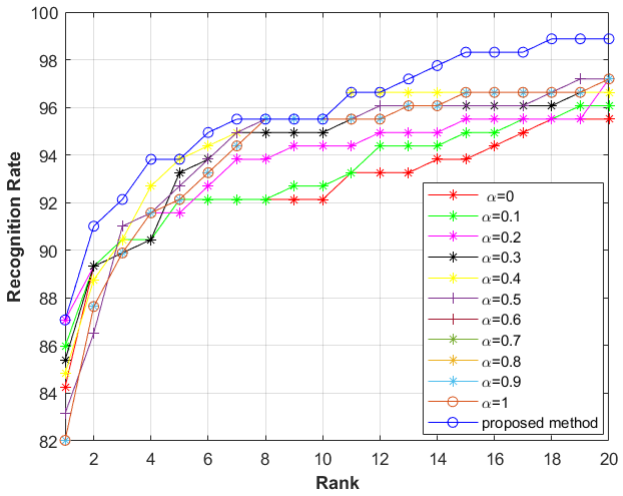

Fig. 9 Comparison of the proposed method with different values of α, in terms of the rank CMC (%), using the PRID dataset.

work [25] that has exploited the periodicity exhibited by a walking person to generate a spatio-temporal body-action model for a video sequence. We have also compared the suggested method with the work of [17], where a signature representation has been produced by modeling a person's appearance as a multi-channel appearance mixture, and each channel is corresponding to a particular region of the body. Table 4 reports the top ranked matching rates of the compared methods on the PRID dataset. It is clear that the proposed method achieves the best matching rates *vs.* hand-crafted features at rank 1 with a value of 87.02%, rank 2 with a value of 93.48%, rank 5 with a value of 95% and rank 20 with a value of 98.25%. Concerning the comparison against deep learning methods, recorded results by the proposed method are very close to the obtained ones using deep learning methods. Indeed, we outperform the work of [28] with rank 1 equals to 87.02% *vs.* 70%, rank 5 equals to 93.48% *vs.* 90%, rank 10 equals to 95% *vs.* 95% and rank 20 equals to 98.25% *vs.* 97%. Table 5 reports the top ranked matching rates of the compared methods on the iLIDS-VID dataset. The proposed method achieves the best matching rates at rank 1 with the value 58.6%, rank 2 with the value 76.26%, rank 5 with the value 82.06% and rank 20 with the value 87.06%.

5 Discussion

The realized experiments show the advantages of the proposed re-identification model over many relevant methods (hand-crafted methods as well as deep learning methods) from the state-of-the-art, even if in some cases it does not

Table 4 Comparison of the proposed method against relevant methods from the state-of-the-art, on the PRID-2011 dataset, in terms of Rank CMC (%).

Hand-crafted				
CMC Rank	1	5	10	20
Wang et al. [42]	40	71.7	84.5	92.2
Liu et al. [25]	64.1	87.3	89.9	92
Khan et al. [17]	70.6	90.2	94.6	97.1
PM	87.02	93.48	95	98.25
Deep Learning				
CMC Rank	1	5	10	20
McLaughlin et al. [28]	70	90	95	97
Chen et al. [6]	93	99.3	100	100

Table 5 Comparison of the proposed method against relevant methods from the state-of-the-art, on the iLIDS-VID dataset, in terms of Rank CMC (%).

Hand-crafted				
CMC Rank	1	5	10	20
Wang et al. [42]	39.5	61.1	71.7	81
Liu et al. [25]	44.3	71.7	83.7	91.7
Khan et al. [17]	33.3	57.8	68.5	80.5
PM	58.6	76.26	82.06	87.66
Deep Learning				
CMC Rank	1	5	10	20
McLaughlin et al. [28]	58	84	91	96
Chen et al. [6]	85.4	96.7	98.8	99.5

reach the best value due to the small used training dataset and/or to the problem of estimating accurately the landmarks, notably for RGB data. For instance, Fig. 10 shows some errors made by the landmark detection algorithm of [5] for some examples within the PRID dataset. On the other side, adopting a limited number of human body joints extracted from RGB sequences, in order to analyze body shape and motion in re-id scenarios, has the major advantage of being independent of appearance, pose and lighting changes. Fig. 11 shows some situations where the gesture and the appearance sharply change within the PRID and the iLIDS-VID datasets. In fact, the skeleton can optimize the manipulation of the data in this case by describing humans using $3D$ coordinates of key body joints. Compared with RGB or depth data, skeleton has many merits like better robustness and much smaller data size. Indeed, the proposed re-identification method has a low computational complexity, and consequently a low processing time, which can be very advantageous in many real-time applications. More precisely, Algorithm 1 is of complexity $O(N.M)$ (Table 6) and the total CPU time of the proposed method (*i.e.* Algorithm 2 including training and testing phases) varies from 16 to 22 minutes, according to the investigated dataset (Table 7). It is worth noting that all the experiments were run using Matlab on a Windows 10 PC with an Intel i7 CPU at 2.20 GHz and 16.00 GB RAM (the source code of the proposed method is available at https://github.com/Elaoud/re-id). Thus, the design of handcrafted features

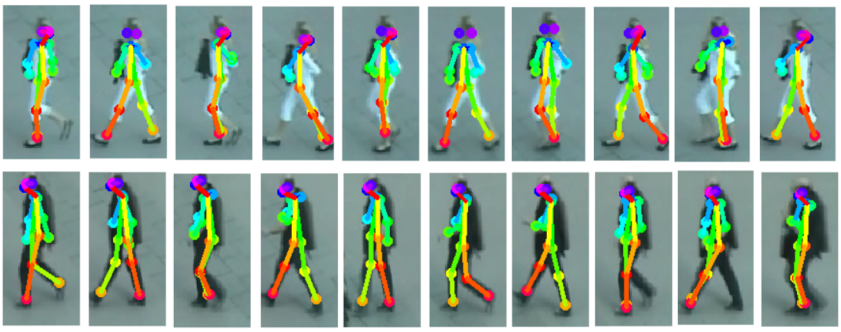

Fig. 10 An illustration of some failure cases of the $2D$ landmark estimation by the algorithm of [5] for some RGB examples from the PRID dataset.

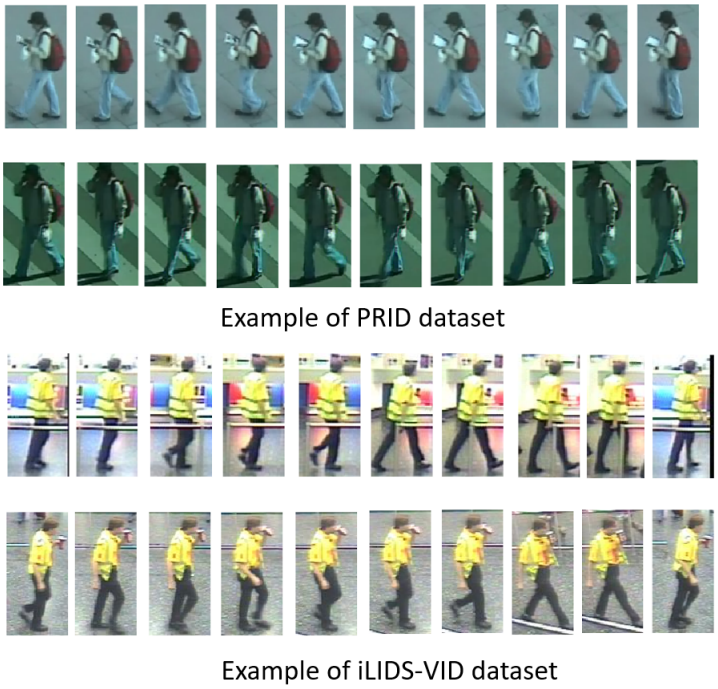

Fig. 11 Examples of sequences, within the studied RGB datasets (PRID and the iLIDS-VID), which show considerable gesture and appearance changes.

allowed us to involve the right trade-off between accuracy and computational efficiency. However, the deep learning methods need large volumes of data, while in most of cases there are lack of annotated data illustrating appearance changes, what can provoke problems in terms of storage and execution cost.

Table 6 Complexity of Algorithm 1 "Weighted linear combination of distances" (the first six row contain the complexity of main blocks of the algorithm and the last row shows the total complexity of the algorithm.

Line(s)	Complexity
7 to 9	$N\ (SizeOf(S1))$
10 to 12	$M\ (SizeOf(S2))$
13	$N.M$
14 to 16	$N.19$
17 to 19	$M.M$
20	$N.M$
$O(N.M)$	

Table 7 CPU time of Algorithm 2 "Dynamic linear combination of distances" for various tested datasets.

Dataset	CPU Time
IAS-Lab RGBD-ID ("TestingA")	20 min
IAS-Lab RGBD-ID ("TestingB")	22 min
BIWI-Lab RGBD-ID ("Walking")	16 min
iLIDS-VID	20 min
PRID 2011	20 min

6 Conclusion

This paper proposes a re-identification method based on hand-crafted features while performing dynamic distance-based comparison. We are interested in modeling and analysing human motion in different views from multiple cameras with 3D joints. We use the skeleton information in order to be independent of clothes', pose and illumination changes. The main contribution of the proposed method resides in evaluating dynamically, according to the input sequence, the similarity between trajectories for re-id using distance on the Grassmann manifold (shape) as well as local distance (anthropometric measures). Indeed, the projection on the Grassmann manifold allows to facilitate the manipulation of sequences of motions while being invariant to rotation. Moreover, we exploit the local distances in order to ensure accurate modeling of individual motions of body parts and anthropometric during the process of re-identifying a person. While using two distances illustrating different information of shape and motion, we try to find the best linear combination with the use of a machine learning technique (random forest) that predicts the best value of the weight α for each person, even with different camera views. In fact, we focus on the best value of α for each person in different views. By adopting this concept of distance learning within the framework of person re-identification, we extract a model that defines a weight dynamically. The predicted weights proved to be able to accurately re-identify persons, even after considerable change in appearance and pose. To improve the current results, we are interested in combining the deep learning techniques along with the used distances in our future work. We can also combine appearance

and skeleton data for more details and precision while dealing with person re-identification.

References

1. Bai, S., Tang, P., Torr, P.H., Latecki, L.J.: Re-ranking via metric fusion for object retrieval and person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 740–749 (2019)
2. Barbosa, I.B., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: European Conference on Computer Vision, pp. 433–442. Springer (2012)
3. Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., Ilic, S.: 3d pictorial structures for multiple human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1669–1676 (2014)
4. Belmonte-Hernández, A., Solachidis, V., Theodoridis, T., Hernandez-Penaloza, G., Conti, G., Vretos, N., Alvarez, F., Daras, P.: Person tracking association using multimodal systems. In: 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2017)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299 (2017)
6. Chen, D., Li, H., Xiao, T., Yi, S., Wang, X.: Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1169–1178 (2018)
7. Chen, Y., He, F., Li, H., Zhang, D., Wu, Y.: A full migration bbo algorithm with enhanced population quality bounds for multimodal biomedical image registration. Applied Soft Computing p. 106335 (2020)
8. Elaoud, A., Barhoumi, W., Drira, H., Zagrouba, E.: Analysis of skeletal shape trajectories for person re-identification. In: International Conference on Advanced Concepts for Intelligent Vision Systems, pp. 138–149. Springer (2017)
9. Fan, X., Jiang, W., Luo, H., Fei, M.: Spherereid: Deep hypersphere manifold embedding for person re-identification. Journal of Visual Communication and Image Representation **60**, 51–58 (2019)
10. Gao, Z., Zhang, H., Dong, S., Sun, S., Wang, X., Yang, G., Wu, W., Li, S., de Albuquerque, V.H.C.: Salient object detection in the distributed cloud-edge intelligent network. IEEE Network **34**(2), 216–224 (2020)
11. Geng, Y., Liang, R.Z., Li, W., Wang, J., Liang, G., Xu, C., Wang, J.Y.: Learning convolutional neural network to maximize pos@ top performance measure. European Symposium on Artificial Neural Networks(ESANN) (2017)
12. Geng, Y., Zhang, G., Li, W., Gu, Y., Liang, R.Z., Liang, G., Wang, J., Wu, Y., Patil, N., Wang, J.Y.: A novel image tag completion method based on convolutional neural transformation. In: International conference on artificial neural networks, pp. 539–546. Springer (2017)
13. Hirzer, M., Beleznai, C., Roth, P.M., Bischof, H.: Person re-identification by descriptive and discriminative classification. In: Scandinavian conference on Image analysis, pp. 91–102. Springer (2011)
14. Imani, Z., Soltanizadeh, H.: Person reidentification using local pattern descriptors and anthropometric measures from videos of kinect sensor. IEEE Sensors Journal **16**, 6227–6238 (2016)
15. Imani, Z., Soltanizadeh, H., Orouji, A.A.: Short-term person re-identification using rgb, depth and skeleton information of rgb-d sensors. Iranian Journal of Science and Technology, Transactions of Electrical Engineering pp. 1–13 (2019)
16. Jing, Y., Si, C., Wang, J., Wang, W., Wang, L., Tan, T.: Cascade attention network for person search: Both image and text-image similarity selection. CoRR **abs/1809.08440** (2018)

17. Khan, F.M., Brèmond, F.: Multi-shot person re-identification using part appearance mixture. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 605–614. IEEE (2017)
18. Li, H., He, F., Chen, Y., Luo, J.: Multi-objective self-organizing optimization for constrained sparse array synthesis. Swarm and Evolutionary Computation **58**, 100743 (2020)
19. Li, H., He, F., Liang, Y., Quan, Q.: A dividing-based many-objective evolutionary algorithm for large-scale feature selection. Soft Computing pp. 1–20 (2019)
20. Li, J., Wang, J., Tian, Q., Gao, W., Zhang, S.: Global-local temporal representations for video person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3958–3967 (2019)
21. Li, W., Zhao, R., Xiao, T., Wang, X.: 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)
22. Li, Y.J., Luo, Z., Weng, X., Kitani, K.M.: Learning shape representations for clothing variations in person re-identification. CoRR **abs/2003.07340** (2020)
23. Liang, G., Lan, X., Chen, X., Zheng, K., Wang, S., Zheng, N.: Cross-view person identification based on confidence-weighted human pose matching. IEEE Transactions on Image Processing **28**(8), 3821–3835 (2019)
24. Liu, H., Jie, Z., Jayashree, K., Qi, M., Jiang, J., Yan, S., Feng, J.: Video-based person re-identification with accumulative motion context. IEEE transactions on circuits and systems for video technology **28**(10), 2788–2802 (2017)
25. Liu, K., Ma, B., Zhang, W., Huang, R.: A spatio-temporal appearance representation for viceo-based pedestrian re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3810–3818 (2015)
26. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2640–2649 (2017)
27. Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y.: Hierarchical gaussian descriptors with application to person re-identification. IEEE transactions on pattern analysis and machine intelligence (2019)
28. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1325–1334 (2016)
29. Munaro, M., Basso, A., Fossati, A., Van Gool, L., Menegatti, E.: 3d reconstruction of freely moving persons for re-identification with a depth sensor. In: 2014 IEEE International Conference on Robotics and Automation (ICRA), pp. 4512–4519. IEEE (2014)
30. Munaro, M., Fossati, A., Basso, A., Menegatti, E., Van Gool, L.: One-shot person re-identification with a consumer depth camera. In: Person Re-Identification, pp. 161–181. Springer (2014)
31. Niu, K., Huang, Y., Ouyang, W., Wang, L.: Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Transactions on Image Processing **29**, 5542–5556 (2020)
32. Paisitkriangkrai, S., Shen, C., Van Den Hengel, A.: Learning to rank in person re-identification with metric ensembles. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1846–1855 (2015)
33. Pala, P., Seidenari, L., Berretti, S., Del Bimbo, A.: Enhanced skeleton and face 3d data for person re-identification from depth cameras. Computers & Graphics **79**, 69–80 (2019)
34. Rao, H., Wang, S., Hu, X., Tan, M., Da, H., Cheng, J., Hu, B.: Self-supervised gait encoding with locality-aware attention for person re-identification. In: Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (2020)
35. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing **26**(1), 43–49 (1978)
36. Singh, A., Kiran, G., Harsh, O., Kumar, R., Singh Rajput, K., SS Vamsi, C., et al.: Real-time aerial suspicious analysis (asana) system for the identification and re-identification

of suspicious individuals using the bayesian scatternet hybrid (bsh) network. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 0–0 (2019)
37. Takač, B., Catala, A., Rauterberg, M., Chen, W.: People identification for domestic non-overlapping rgb-d camera networks. In: 2014 IEEE 11th International Multi-Conference on Systems, Signals & Devices (SSD14), pp. 1–6. IEEE (2014)
38. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Advances in neural information processing systems, pp. 1799–1807 (2014)
39. Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. International Journal of Computer Vision pp. 1–47 (2019)
40. Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L.: Joint learning of single-image and cross-image representations for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1288–1296 (2016)
41. Wang, J., Zhong, Y., Li, Y., Zhang, C., Wei, Y.: Re-identification supervised texture generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11846–11856 (2019)
42. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: European conference on computer vision, pp. 688–703. Springer (2014)
43. Wang, Z., Wei, D., Hu, X., Luo, Y.: Human skeleton mutual learning for person re-identification. Neurocomputing (2020)
44. Wu, A., Zheng, W.S., Gong, S., Lai, J.: Rgb-ir person re-identification by cross-modality similarity preservation. International Journal of Computer Vision pp. 1–21 (2020)
45. Wu, Y., Bourahla, O.E.F., Li, X., Wu, F., Tian, Q.: Adaptive graph representation learning for video person re-identification. IEEE Transactions on Image Processing (2020)
46. Yang, Y., Lei, Z., Wang, J., Li, S.Z.: In defense of color names for small-scale person re-identification. In: 2019 International Conference on Biometrics (ICB), pp. 1–6. IEEE (2019)
47. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3d pose estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4948–4956 (2016)
48. Yasuko, C.: Statistics on special manifolds, lecture notes in statistics. in vol. 174 (2003)
49. Zhang, G., Liang, G., Li, W., Fang, J., Wang, J., Geng, Y., Wang, J.Y.: Learning convolutional ranking-score function by query preference regularization. In: International conference on intelligent data engineering and automated learning, pp. 1–8. Springer (2017)
50. Zhang, G., Liang, G., Su, F., Qu, F., Wang, J.Y.: Cross-domain attribute representation based on convolutional neural network. In: International Conference on Intelligent Computing, pp. 134–142. Springer (2018)
51. Zhang, L., Shi, Z., Zhou, J.T., Cheng, M.M., Liu, Y., Bian, J.W., Zeng, Z., Shen, C.: Ordered or orderless: A revisit for video based person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
52. Zhang, L., Zhang, L., Du, B., You, J., Tao, D.: Hyperspectral image unsupervised classification by robust manifold matrix factorization. Information Sciences **485**, 154–169 (2019)
53. Zhang, R., Li, J., Sun, H., Ge, Y., Luo, P., Wang, X., Lin, L.: Scan: Self-and-collaborative attention network for video person re-identification. IEEE Transactions on Image Processing **28**(10), 4870–4882 (2019)
54. Zhao, C., Wang, X., Zuo, W., Shen, F., Shao, L., Miao, D.: Similarity learning with joint transfer constraints for person re-identification. Pattern Recognition **97**, 107014 (2020)
55. Zheng, W., Li, L., Zhang, Z., Huang, Y., Wang, L.: Relational network for skeleton-based action recognition. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 826–831. IEEE (2019)