Analysis of Skeletal Shape Trajectories for Person Re-Identification

Amani Elaoud, Walid Barhoumi, Hassen Drira, Ezzeddine Zagrouba

Analysis of Skeletal Shape Trajectories for Person Re-Identification

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

[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, 45 Rue des Entrepreneurs, 2035 Tunis-Carthage, Tunisia
[3] IMT Lille Douai, CRISTAL (UMR CNRS 9189), Villeneuve-d'Ascq, France
amani89elaoudz@gmail.com, walid.barhoumi@enicarthage.rnu.tn, hassen.drira@imt-lille-douai.fr, ezzeddine.zagrouba@fsm.rnu.tn

Abstract. In this paper, we are interested in people re-identification using skeleton information provided by a consumer RGB-D sensor. We perform the modelling and the analysis of human motion by focusing on 3D human joints given by skeletons. In fact, the motion dynamic is modeled by projecting skeleton information on Grassmann manifold. Moreover, in order to define the identity of a test trajectory, we compare it against a labeled trajectory database while using an unsupervised similarity assessment procedure. Indeed, the main contribution of this work resides in the introduced distance that combines temporal information as well as global and local geometrical ones. Realized experiments on standard datasets prove that the proposed method performs accurately even though it does not assume any prior knowledge.

Keywords: Person re-identification, RGB-D sensors, Skeleton information, Grassmann manifold, Similarity evaluation

1 Introduction

Recognizing the different persons in different locations is considered as an important and challenging task. Modeling such long-term activities has a great interest and was the subject of several applications in different fields like service robotics [14], smart environments [9] and video surveillance [3]. An emerging solution is to identify people through the analysis of the way of walk (often referred to as gait). In fact, analysis of walk is not limited to medical or sports and can play a decisive role in security and surveillance as behavioral biometric factor. It can afford effective solutions for many environments, such as airports, museums and shopping malls [2] since it does not require individuals to carry scanning device, waiting in line or remembering user passwords. According to Prakash et al. [15], walking is considered as one of the natural and common

traits for the human being, but from the perspective of analysis, it is among one of the most complex phenomena. Thus, re-identification (re-id) is a very interesting research topic in computer vision, which represents a fundamental issue. Methods used for re-id based on gait analysis can be divided into two categories: with markers [16] and others without markers [17]. We are interested herein in re-id of persons by gait analysis without using markers. In this case, re-id methods can be coarsely grouped, according to the used data, into three classes: RGB appearance-based re-id, depth-based re-id and skeleton-based re-id. Most existing works rely on RGB-based appearance features. For example, Harandi et al. have demonstrated the importance of RGB information with Riemannian locality preserving projection in order to identify individuals. In a study of Alavi et al. [1], a metric learning is proposed in order to automatically re-identify a given person among a large set of candidates over various non-overlapping camera views. More recently, Zaho et al. [20] used patch-based local features to estimate the similarity between two images. In fact, filters were used for person re-id while combining SIFT with color histogram as features. Nevertheless, RGB sources suffer always from pose changes, occlusions, clothes change and illumination variations. Generally, re-id approaches based on appearance models from RGB videos, assume that individuals do not change their clothing within the observation period [14] [1]. This hypothesis represents a very strong restriction, since it constraints re-id methods to be applied under a limited temporal range (reasonably, in the order of minutes). In fact, the persons appearance is subject to dramatic changes in camera views due to changes in the view angle, the body pose, the illumination and the background clutter. For instance, the same person can have different images depending on the different clothes he or she wears or the illumination conditions. With the recent advances in 3D imaging devices, such as Microsoft Kinect, and given the above mentioned drawbacks of RGB-based methods, there is a growing interest in depth-based and skeleton-based approaches. Indeed, these informations are robust to variations of illumination, scale and rotation [8]. On the one hand, depth-based methods aim to create human signature from depth images or videos in the absence of RGB information [10]. The used models are volumetric and can leverage 4D inputs to be robust against appearances and changes. On the other hand, human representations based on 3D skeleton information have been intensively studied. The main advantage resides in the fact that a small number of joint positions can effectively represents human motion. Thanks to the work of Shotton et al. [18], the skeleton can be extracted in real-time from depth data. Consequently, skeleton-based information was exploited and kept attracting an increasing attention for person re-id. The study of Murano et al. [14] proposed skeletal tracker to determine joints as keypoints. They extract 2D and 3D descriptors in order to compute signatures of people. By comparing signatures of each target in the test frame with those in the training frames, the best match is selected as a result. Barbosa et al. [2] proposed also a person signature that exploits skeleton-based features while exploiting soft-biometric features using a set of ratios of joint distances. Moreover, independently of the used data, existing re-id methods can develop a

supervised or an unsupervised way to identify persons. Both strategies have their own advantages. While the first group of approaches is more flexible regarding data, feature quality and selection, the second approach can work autonomously, without any need to training data. In this paper, we present an unsupervised method for person re-id based on similarity assessment within skeleton information in order to be able to provide superior performance with appearance changes. In fact, skeleton-based approaches have a strong impact on a number of applications but they are not much used in the field of people re-id. The main idea of this work is the projection of a set of videos, presented by 3D human joints, on Grassmann manifold. In fact, the Grassmann manifold is a convenient representation to deal with unsupervised re-id of individuals in diverse locations. Several modern applications explored this manifold, such as face recognition, object recognition and action recognition. For example, harandi et al. [11] define a convenient way of dealing with image sets in order to represent them as points on Grassmann manifold. Gopalan et al. [6] have demonstrated a promising approach to tackle unsupervised object recognition by connecting the source subspace and the target subspace on the Grassmann manifold. More recently, body joints were represented as trajectories on the manifold to obtain an action-recognition system [19]. In this work, we are interested in this manifold to perform people re-id based on RGB-D dataset. We use a richer representation for exploiting as much space-time information from skeletons projected in Grassmann manifold. Besides, in order to identify a person within a test video, we assess the similarity between trajectories while introducing an efficient distance that combines global and local features.
The rest of this paper is organized as follows. Section 2 presents the proposed unsupervised method for re-identifying individuals. Experimental results are discussed in Section 3. Section 4 concludes the paper and provides some future ideas.

2 Proposed method

The pipeline of the proposed re-id method is presented in Fig. 1. In fact, given a test skeleton sequence V_t, composed of a set of n_t ordered frames that show a person S_t in motion, the goal is to identify this person while exploring a labeled RGB-D video gallery set. This gallery set contains K videos V_i ($i \in \{1, \ldots, K\}$), which are composed of different frame numbers n_i ($i \in \{1, \ldots, K\}$). The K persons in the gallery set, as well as the input one S_t, are performing the same motion. The proposed method is based only on skeleton data within the 3D information while ignoring RGB channel, in order to ensure robustness against complex background and to be insensitive to appearance and lightning changes [8].

Thus, each frame in a skeleton video of the person S_i is modeled by a matrix M_i^l of size 20×3 (the (x, y, z) coordinates of the 20 joints), where $l \in \{1, \ldots, n_i\}$ and n_i denotes the number of frames in the studied video V_i. Such a matrix is then projected on the Grassmann manifold using Singular Value Decompo-

sition (SVD)-based orthogonalization. Hence, the skeleton sequences are modeled as trajectories on Grassmann manifold. The projection of skeleton data on Grassmann manifold is motivated by the invariance against to rotation, and the skeleton-based human representation is used in order to ensure robustness against variations of illuminations and human body scale as it is estimated from 3D data. Given an unlabeled test video V_t, the same processing is applied in order to obtain a trajectory T_t, illustrating the motion for the test video. Lastly, a similarity evaluation is applied to identify the person in the input video V_t by comparing its trajectory T_t against the gallery ones T_i ($i \in \{1, \ldots, K\}$).

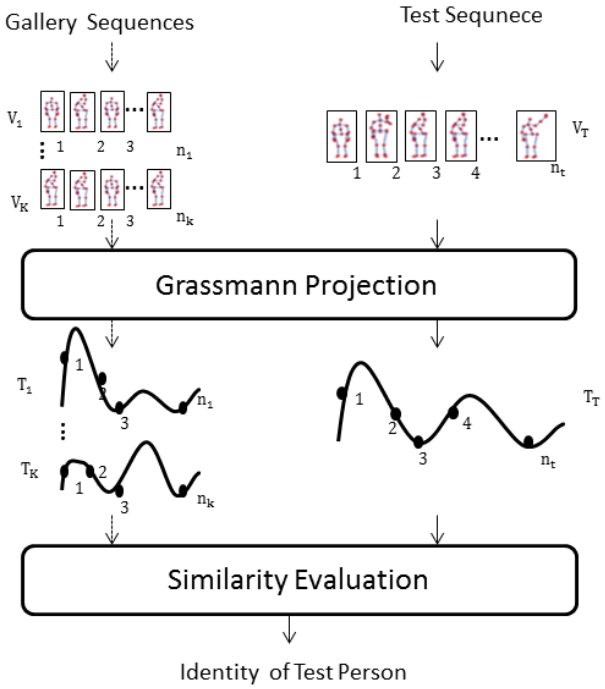


Fig. 1. Outline of the proposed re-identification method. The skeleton sequences are modeled as trajectories on Grassmann manifold, where the similarity evaluation is then performed.

2.1 Projection on Grassmann manifold

A manifold is a topological space that is locally similar to Euclidean space. Manifold analysis has been extensively considered with success by various works including skeleton analysis such as the work of Slama et al. [19]. The main

drawback of this work is the linear approximation by using ARMA model (autoregressive moving average model) for time modeling. Moreover, elements of Grassmann manifold model sequences are not skeletons and the projection on Grassmann manifold is not reversible as only the k-first principal components are kept. In our case, the projection on Grassmann manifold is without any loos of information. Actually, each skeleton is projected on Stiefel manifold and the distance is then calculated on the Grassmann manifold which represents in our case the quotient space of Steifel when the group of rotation in R^3 is considered. More formally, Stiefel manifold is the set of k-dimensional orthonormal bases in R^n where $k < n$, represented by $\mathrm{V}(R^n)$. Thus, a point on a Stiefel manifold is an orthonormal matrix of size 20×3 that can be viewed as 3-dimensional subspace of R^{20}. Grassmann manifold $\mathrm{G}(R^n)$ is a quotient space of $\mathrm{V}(R^n)$, represented by $\mathrm{V}(R^n)/\mathrm{SO(K)}$, where SO(K) is the orthogonal group of dimension k. Two points on a Grassmann manifold are equivalent if one can be mapped into the other one by a rotation matrix: a 3×3 orthogonal matrix [4]. Given a skeleton sequence, the proposed approach suggests to model each frame as a point in Grassmann manifold (a matrix). This matrix is represented by 3D coordinates of body joints denoting the spatial position of the 20 joints. Thus, the video is transformed on a trajectory that links different points on Grassmann manifold. The projection on Grassmann manifold requires an SVD-based orthogonalization. The main advantages of this projection is being reversible and has no loss of information. The next step consists on similarity computing between human skeletal joint trajectories in order to identify the identity of a given skeleton sequence.

2.2 Similarity evaluation

The proposed method tries to re-identify a person by the simple way using a measure of similarity. In fact, we evaluate the dissimilarity between gallery sequences and the test one in order to find the shortest value of dissimilarity. To do this, we firstly use Dynamic Time Warping (DTW) [5] which is a popular sequence alignment and matching method especially for two unregulated sequences. DTW algorithm resolves the problem of temporal alignment and measures the similarity between sequences varied in time. In this paper, the matching was performed while using two different measures of dissimilarity: Euclidean dissimilarity (1) and Grassmann dissimilarity (2). The first dissimilarity is calculated without the use of any manifold, according to:

$$Diss_{Euclidean}(E_t, E_i) = \sum_{l=1}^{n_t} \|(E_t^*(l) - E_i^*(l))\| = \sum_{l=1}^{n_t} \left\|(M_t^l - M_i^l)\right\|, \quad (1)$$

where, E_t and E_i denote the trajectories of the persons S_t and S_i, respectively, and E_t^* and E_i^* are the warped trajectories of E_t and E_i after the alignment based on DTW.

Furthermore, the Geodesic distance is the length of the shortest curve connecting P_t^l and P_i^l which are the projection of the given two points M_t^l and M_i^l on the Grassmann manifold and it is computed according to:

$$Diss_{Grassmann}(T_t, T_i) = \sum_{l=1}^{n_t} Diss_{Grassmann}(T_t^*(l), T_i^*(l)) = \sum_{l=1}^{n_t} \theta_l^2, \quad (2)$$

where T_t^* and T_i^* are the warped trajectories (aligned via dynamic time warping of T_t and T_i in Grassmann manifold) and θ_l is the principal angle between the subspaces generated by P_t^l and P_i^l. Fig. 2 illustrates the calculation of principal angles between subspaces in Grassmann manifold.

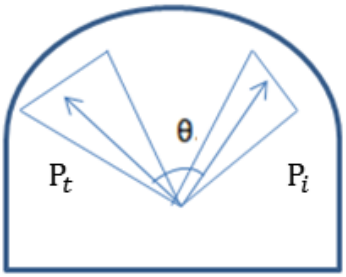


Fig. 2. Illustration of principal angles calculation on Grassmann manifold.

Besides, we compute distances between adjacent joints which correspond to parts lengths. These distances are characteristic for each person and are used as local features, in addition to Euclidean and Grassmann distances between corresponding frames along the two aligned sequences. The combination of local features (denoted by local distance) and Grassmann distance are then combinated to get the dissmilarity between corresponding frames. This dissimilarity between a test sequence aligned to a gallery one is computed as the sum of the dissimilarity between corresponding frames while the closest neighbor is used for classification.

3 Experiments

Despite of the increasing interest in performing person re-id using 3D data, the majority of human-based RGB-D datasets were constructed to validate and to assess human activity analysis and action recognition methods. Indeed, these datasets are not suited for the evaluation of re-id methods. Among the very few publicly available re-id benchmarks, the most used one is BIWI-Lab RGBD-ID

(Munaro et al. [12]) which is widely used for long-term people re-id from RGB-D cameras. It illustrates individuals performing a certain routine of motions and walks in front of a Kinect sensor. There is also the challenging IAS-Lab RGBD-ID dataset proposed by Munaro et al. [12]. It contains different sequences of people with two sets of testing. The first set of sequences (named "TestingA") was performed with people wearing different clothes compared to training sequences, while the second one (named "TestingB") was collected in a different room, but with the same clothes (Fig. 3). Firstly, we start by testing the proposed method on the IAS-Lab RGBD-ID dataset, since the same person repeats the same action in the training set as well as in the testing set. Thus, it can be beneficial to detect a person by his motion (*i.e.* the way of walking). This standard dataset consists of 33 sequences of 11 different people, where 11 sequences compose the training set and the remaining 22 sequences form the testing set (Fig. 3).

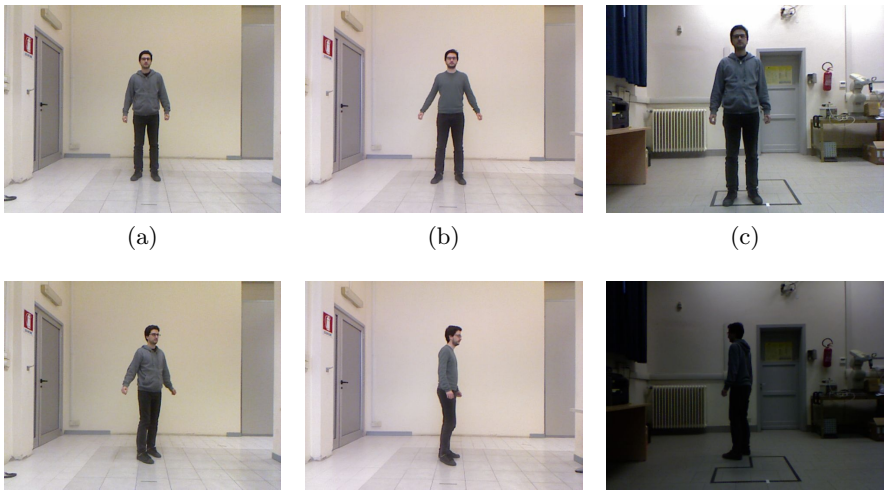


Fig. 3. A sample of IAS-Lab dataset: (a)-(c) shows the appearance changes (d)-(f) shows motions with rotation.

Secondly, we validated the suggested method on the BIWI-Lab RGBD-ID dataset [12]. This dataset contains 50 training sequences and 56 testing sequences of 50 different people, performing a certain routine of motions in front of a Kinect sensor. For every person in the testing set, a "Still" sequence and a "Walking" sequence have been collected. For evaluation purposes, we compute Cumulative Matching Curves (CMC) [7], which are commonly used for analyzing the re-id performances. These curves present the cumulative identification rates as a function of the rank distribution. It can therefore be said that the more the rank increases, the more the corresponding recognition rate is linked to a lower level. In Table 1, we report the rank-1 for the IAS-Lab RGBD-ID dataset (for "TestingA"

set as well as for "TestingB" set) and for the BIWI-Lab RGBD-ID dataset (for "Walking" set as well as for "Still" set). We show that the CMC obtained with local distance gives the best results such as it reaches the maximum value of 63.63% with the set "TestingB" of the IAS-Lab RGBD-ID dataset. As it can be noticed, for the global distances, the Geodesic one outperforms the Euclidean distance, expect for the "TestingA" set of the IAS-Lab RGBD-ID dataset. For instance, for the set "TestingB" of the IAS-Lab RGBD-ID dataset, Geodesic distance records 45.45% as rank-1 against 36.36% with Euclidean distance. The importance of the Geodesic distance can be explained by the fact that this distance is invariant to rotation. Indeed, we recorded a null Geodesic distance between a skeleton of a person and its skeleton after performing several rotations.

Table 1. Rank-1 with the testing sets of IAS-Lab RGBD-ID and BIWI-Lab RGBD-ID datasets.

Testing set	Accuracy (%)
Global distance (Geodesic distance)	
TestingA	27.27%
TestingB	45.45%
Walking	7.14%
Still	28.57%
Global distance (Euclidean distance)	
TestingA	36.36%
TestingB	36.36%
Walking	7.14%
Still	25%
Local distance	
TestingA	45.45%
TestingB	63.63%
Walking	10.71%
Still	28.57%

In the following, we will show that the combination of heterogeneous distances, among the best ones (Geodesic distance + Local distance), will further enhance the person re-id accuracy. In fact, we compared the results reported in [12], that are obtained with a relevant supervised method based on Neural Networks (NN), with our results (Figure 4) on the IAS-Lab RGBD-ID dataset using its two test sets: "TestingA" (Fig. 4.a) and "TestingB" (Fig. 4.b). We see that the recognition rate using the proposed method is better than [12] at all the ranks. For example, the rank-1 gives significant value compared to [12] (45.45% for "TestingA" of IAS-Lab against 25.5%, and 72.72% for "TestingB" of IAS-Lab against 62%). Moreover, the rank-2 recognition rate using "TestingA" of IAS-Lab is 73% for the proposed method against 42% for [12] (Fig. 4.a). For the

"TestingB" of IAS-Lab, we can also notice that the rank-2 recognition records 92% for the proposed method and only 73% for [12] (Fig. 4.b).

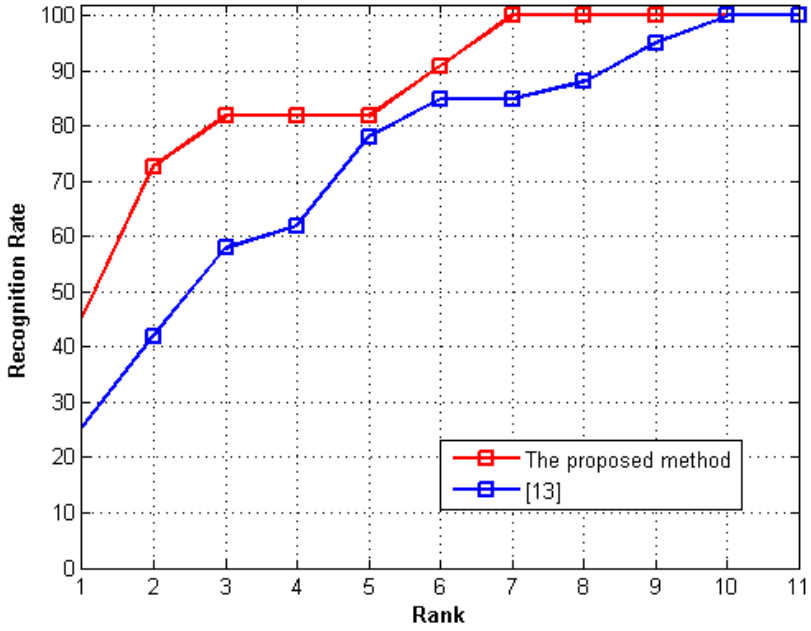
Besides, we compared our method, using the BIWI-Lab RGBD-ID dataset, with two skeleton-based methods: [13], which is a supervised method based on Support Vector Machine (SVM), and [12] (Fig. 5). As reported in Fig. 5.b, we show better performance at rank-1 (28.57% for the proposed method against 26.6% for [12] and 11.6% for [13]) using the test set "Still" of BIWI-Lab RGBD-ID. However, the remaining ranks are similar for the three methods. As the test set "Walking" of BIWI-Lab RGBD-ID is concerned (Fig. 5.a), the proposed method performs similar results at all the ranks compared to [12] and [13]. Thus, we can conclude that the proposed unsupervised re-id method significantly outperforms state-of-the-art methods even if, it does not assume any prior knowledge, notably for IAS-Lab dataset and for Still skeletons in BIWI-Lab RGBD-ID dataset. We believe that this is mainly due to the invariance to rotation thanks to Grassmann projection. Actually, the datasets involve several motions with rotation such as rotation around the vertical axis.

4 Conclusions and future works

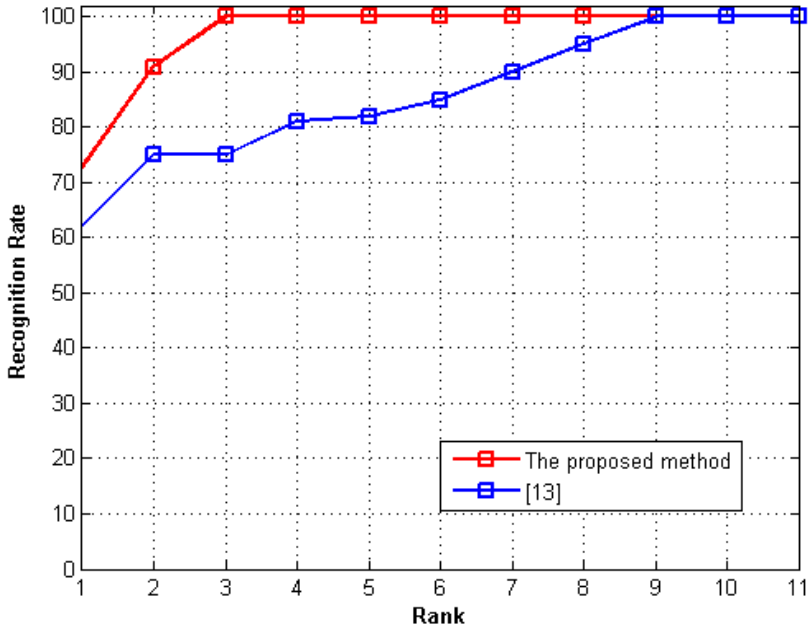
In this work, we proposed to perform the modeling and the analysis of human motion by focusing only on 3D human joints given by skeletons. To overcome the problem of the different poses a person can have, we exploited the skeletal information provided by a low cost RGB-D sensor. The dynamic of motion is modeled by projecting skeleton information on Grassmann manifold. We showed how this modeling can be effectively used for the long-term re-id task while introducing a distance that combines temporal as well as global and local geometrical information. The classification is simply performed by closest neighbor method. Machine learning techniques will be investigated in the future to improve the classification performance.

References

1. Alavi, A., Yang, Y., Harandi, M., Sanderson, C.: Multi-shot person re-identification via relational stein divergence. In: Image Processing (ICIP), 2013 20th IEEE International Conference on. pp. 3542–3546. IEEE (2013)
2. Barbosa, I., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: Computer Vision–ECCV 2012. Workshops and Demonstrations. pp. 433–442. Springer (2012)
3. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: Pattern Recognition (ICPR), 2010 20th International Conference on. pp. 1413–1416. IEEE (2010)
4. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM journal on Matrix Analysis and Applications 20(2), 303–353 (1998)

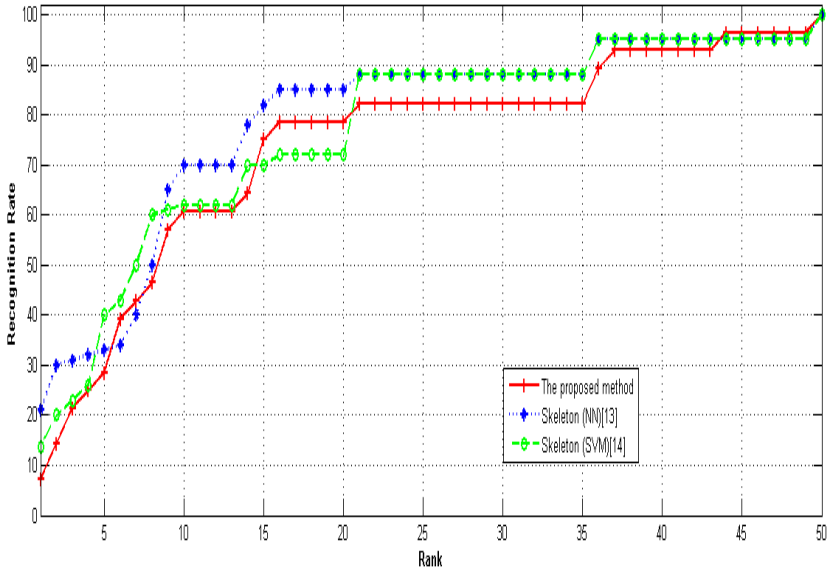


(a)

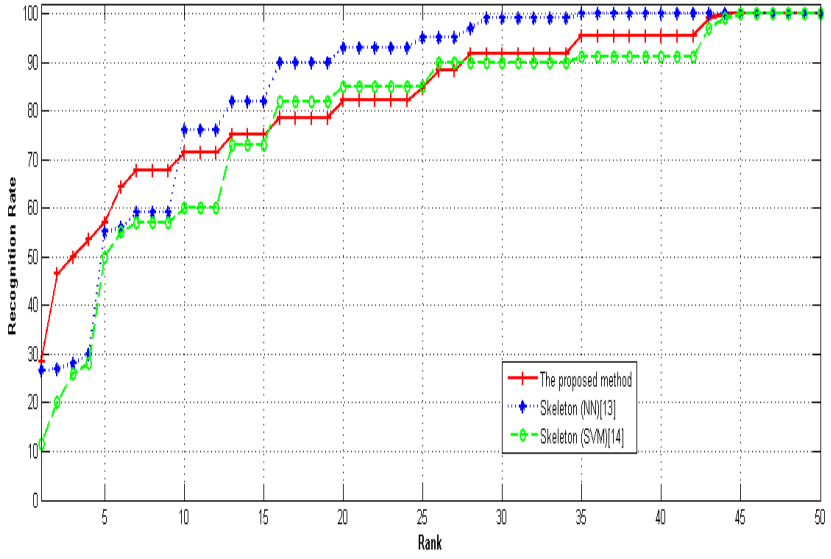


(b)

Fig. 4. Comparison of the recognition rates of the proposed method against [12] while using the two testing sets of the IAS-Lab RGBD-ID dataset ((a) "TestingA" and (b) "TestingB")



(a)



(b)

Fig. 5. Comparison of the proposed method against [12] and [13] using (a) "Walking" and (b) "Still" testing sets of the BIWI-Lab RGBD-ID dataset.

5. Giorgino, T., et al.: Computing and visualizing dynamic time warping alignments in r: the dtw package. Journal of statistical Software 31(7), 1–24 (2009)

6. Gopalan, R., Li, R., Chellappa, R.: Domain adaptation for object recognition: An unsupervised approach. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 999–1006. IEEE (2011)
7. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. Computer Vision–ECCV 2008 pp. 262–275 (2008)
8. Han, F., Reily, B., Hoff, W., Zhang, H.: Space-time representation of people based on 3d skeletal data: A review. Computer Vision and Image Understanding (2017)
9. Han, J., Han, J.: Rgb-d human identification and tracking in a smart environment. In: Computer Vision and Machine Learning with RGB-D Sensors, pp. 195–211. Springer (2014)
10. Haque, A., Alahi, A., Fei-Fei, L.: Recurrent attention models for depth-based person identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1229–1238 (2016)
11. Harandi, M.T., Sanderson, C., Shirazi, S., Lovell, B.C.: Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. pp. 2705–2712. IEEE (2011)
12. 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: Robotics and Automation (ICRA), 2014 IEEE International Conference on. pp. 4512–4519. IEEE (2014)
13. 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)
14. Munaro, M., Ghidoni, S., Dizmen, D.T., Menegatti, E.: A feature-based approach to people re-identification using skeleton keypoints. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. pp. 5644–5651. IEEE (2014)
15. Prakash, C., Kumar, R., Mittal, N.: Recent developments in human gait research: parameters, approaches, applications, machine learning techniques, datasets and challenges. Artificial Intelligence Review pp. 1–40 (2016)
16. Prakash, C., Mittal, A., Kumar, R., Mittal, N.: Identification of spatio-temporal and kinematics parameters for 2-d optical gait analysis system using passive markers. In: Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in. pp. 143–149. IEEE (2015)
17. Prakash, C., Mittal, A., Tripathi, S., Kumar, R., Mittal, N.: A framework for human recognition using a multimodel gait analysis approach. In: International conference on computing, communication and automation (ICCCA2016), IEEE (2016)
18. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM 56(1), 116–124 (2013)
19. Slama, R., Wannous, H., Daoudi, M., Srivastava, A.: Accurate 3d action recognition using learning on the grassmann manifold. Pattern Recognition 48(2), 556–567 (2015)
20. Zhao, R., Ouyang, W., Wang, X.: Learning mid-level filters for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 144–151 (2014)