One-Shot Person Re-Identification with a Consumer Depth Camera

Matteo Munaro, Andrea Fossati, Alberto Basso, Emanuele Menegatti and Luc Van Gool

Abstract In this chapter, we propose a comparison between two techniques for one-shot person re-identification from soft biometric cues. One is based upon a descriptor composed of features provided by a skeleton estimation algorithm; the other compares body shapes in terms of whole point clouds. This second approach relies on a novel technique we propose to warp the subject's point cloud to a standard pose, which allows to disregard the problem of the different poses a person can assume. This technique is also used for composing 3D models which are then used at testing time for matching unseen point clouds. We test the proposed approaches on an existing RGB-D re-identification dataset and on the newly built *BIWI RGBD-ID* dataset. This dataset provides sequences of RGB, depth and skeleton data for 50 people in two different scenarios and it has been made publicly available to foster advancements in this new research branch.

1 Introduction

The task of identifying the person that is in front of a camera has plenty of important practical applications: Access control, video-surveillance, and people tracking are a few examples of such applications.

The computer vision problem that we tackle in this paper is inside the branch of non-invasive and non-cooperative biometrics. This implies not having access to more reliable and discriminative data such as the DNA sequence and fingerprints, but simply relying on the input provided by a cheap consumer depth camera.

Matteo Munaro, Alberto Basso and Emanuele Menegatti
Intelligent Autonomous Systems Laboratory, University of Padua, Via Gradenigo 6a, 35131 - Padua, e-mail: \{munaro,bassoal1,emg\}@dei.unipd.it

Andrea Fossati and Luc Van Gool
Computer Vision Laboratory, ETH Zurich, Sternwartstrasse 7, 8092 Zurich, e-mail: \{fossati,vangool\}@vision.ee.ethz.ch

We decided to take advantage of a depth-sensing device to overcome a few shortcomings intrinsically present in standard video-based re-identification. These include for example non-invariance to different viewpoints and lighting conditions, in addition to being very sensitive to clothing appearance. On the other hand the known disadvantages of consumer depth cameras, i.e. sensitivity to solar infra-red light and limited functioning range, do not usually constitute a problem in standard re-identification scenarios.

The set of features that we adopt to identify a specific person are commonly known as soft biometrics. This means that each feature alone is not a univocal identifier for a certain subject. Still, the combination of several soft biometrics features can show a very good discriminative performance even within large sets of persons.

We take into account both skeleton lengths and the global body shape to be able to describe a subject's identity. Moreover, we extract also facial features for comparison purposes. All the necessary information is collected using a single device, namely a Microsoft Kinect. Given that a body shape can vary also because of the different poses the subject can assume, we warp every point cloud back to a standard pose before comparing them.

Both the approaches we propose in this chapter aim at a one-shot re-identification. After a training phase, during which the classifier parameters or the training models are learned for each of the subjects in the dataset, the system is able to estimate the ID label of detected people separately for each input frame, in real-time. To improve robustness in the estimation, the output of multiple consecutive frames can be easily integrated, for example using a voting scheme.

The contributions of this chapter are three-fold: On one hand we propose a novel technique for exploiting skeleton information to transform persons' point clouds to a standard pose in real-time. Moreover, we explain how to use this transformed point clouds for composing 3D models of moving people which can be used for re-identification by means of an ICP matching with new test clouds and we compare this approach with feature-based approaches which classify skeleton and face descriptors. Finally, we present a novel biometrics RGB-D dataset including 50 subjects: For each subject we provide a sequence including standard video input, depth input, a segmentation mask and the skeleton as provided by the Kinect SDK. Additionally, the dataset includes several labeled testing sequences collected in a different scenario.

2 State of the Art

As cheap depth-sensing devices have started appearing in the market only very recently, the literature in this specific field is quite limited. We will first introduce several vision-based soft biometrics approaches, and then analyze in more detail a few depth-based identification techniques.

The integration of multi-modal cues for person identification is an active research topic since the 90s [14]. For example, in [9] the authors integrate a voice-based

system with face recognition using hyper basis function networks. The concept of information fusion in biometrics has been methodically studied in [22], in which the authors propose several different architectures to combine multiple vision-based modalities: Fusion can happen at the feature extraction level, which usually consists in concatenating the input feature vectors. Otherwise there can be fusion at the matching score level, by combining the scores of the different sub-systems, or at the decision level, i.e. each sub-system takes a decision and then decisions are combined, for example through a majority voting scheme.

Most vision-based systems fall in the category of soft-biometrics, which are defined to be a set of characteristics that provide some biometric information, but are not able to individually authenticate the person, mainly due to lack of distinctiveness and permanence [15].

Vision-based biometrics systems can be either collaborative, as for example iris recognition or fingerprint analysis, or non-collaborative. We will mainly focus on non-collaborative traits, as they are more generally applicable and easier to process: Face-based identification is a deeply studied topic in the computer vision literature [34]. Efforts have been spent in making it more robust to different alignment and illumination conditions [28], and to small training set sizes [35]. The problem has also been tackled in a real-time setup [1, 16] and from a 3D perspective [6]. Another type of vision-based analysis that has been used for people identification is gait recognition [29, 21], which can be either model-based, i.e. a skeleton is first fitted to the data, or model-free, for example by analysing directly silhouettes. This is by definition a soft biometrics, as it is in general not discriminative enough to identify a subject, but can be very powerful if combined with other traits. Finally, also visual techniques have been proposed, that try to re-identify a subject based on a global appearance model [12, 4, 31]. The intrinsic drawback of such approaches is that they can only be applied to tracking scenarios and are not suitable for long time-span recognition.

As mentioned above, due to the very recent availability of cheap depth sensing devices, only a few works exist that focused on identification using such multimodal input. In [19], it is shown that anthropometric measures are discriminative enough to obtain a 97% accuracy on a population of 2000 subjects. The authors apply Linear Discriminant Analysis to very accurate laser scans to obtain such performance. Also the authors of [26] studied a similar problem. They in fact used a few anthropometric features, manually measured on the subjects, as a pre-processing pruning step to make face-based identification more efficient and reliable. In [23], the authors have recently proposed an approach which uses the input provided by a network of Kinect cameras: the depth data in their case is only used for segmentation, while their re-identification techniques purely relies on appearance-based features. The authors of [2] propose a method that relies only on depth data, by extracting a signature for each subject. Such signature includes features extracted from the skeleton as the lengths of a few limbs and the ratios of some of these lengths. In addition, geodesic distances on the body shape between some pairs of body joints are considered. The choice of the most discriminative features is based upon experiments carried on a validation dataset. The signatures are extracted from

a single training frame for each subject, which renders the framework quite prone to noise, and weighted euclidean distances are used to compute distances between signatures. The weights of the different feature channels are simply estimated through an exhaustive grid search. The dataset used in the paper has also been made publicly available, but this does not contain facial information of the subjects, in contrast with the dataset proposed within this paper. Also Kinect Identity [17], the software running on the Kinect for XBox360, uses multi-modal data, namely the subject's height, a face descriptor and a color model of the user's clothing to re-identify a player during a gaming session. In this case, though, the problem is simplified as such re-identification only covers a very short time-span and the number of different identities is usually very limited.

3 Datasets

With the recent availability of cheap depth sensors, a lot of effort in the computer vision community has been put into collecting novel datasets. In particular, several groups have proposed databases of human motions, usually making available skeleton and depth data in conjunction with regular RGB input [33, 18, 30, 25, 32, 20]. Nonetheless, the vast majority of these are focusing on human activity analysis and action recognition, and for this reason they are generally composed by many gestures performed by few subject.

On the other hand, the problem we tackle in this paper is different and requires data relative to many different subjects, while the number of gestures is not crucial. From this perspective, only a dataset has been proposed so far [2]. It consists of 79 different subjects collected in four different scenarios. The collected information, for each subject and for each scenario, includes five RGB frames (in which the face has been blurred), the foreground segmentation mask, the extracted skeleton, the corresponding 3D mesh and an estimation of the ground plane. This dataset contains very few frames for each subject, thus machine learning approaches can be hardly tested because of the little data available for training a person classifier. Moreover, the faces of the recorded subjects have been blurred for privacy reasons, making the comparison with a baseline built upon face recognition impossible.

3.1 BIWI RGBD-ID Dataset

To perform more extensive experiments on a larger amount of data we also collected our own RGB-D Identification dataset called *BIWI RGBD-ID*[1]. It consists of video sequences of 50 different subjects, performing a certain routine of motions in front of a Kinect, such as a rotation around the vertical axis, several head movements

[1] The *BIWI RGBD-ID* dataset can be downloaded at: `http://robotics.dei.unipd.it/reid`.

and two walks towards the camera. The dataset includes synchronized RGB images (captured at the highest resolution possible with the Kinect, i.e. 1280×960 pixels), depth images, persons' segmentation maps and skeletal data (as provided by the Kinect SDK), in addition to the ground plane coordinates. These videos have been acquired at about 10fps and last about one minute for every subject.

Moreover, we have collected 56 testing sequences with 28 subjects already present in the dataset. These have been collected on a different day and therefore most subjects are dressed differently. These sequences are also shot in different locations than the studio room where the training dataset had been collected. For every person in the testing set, a *Still* sequence and a *Walking* sequence have been collected. In the *Walking* video, every person performs two walks frontally and two other walks diagonally with respect to the Kinect.

4 Approach

The framework we have designed allows to identify a subject standing in front of a consumer depth camera, taking into account a single input frame. To achieve this goal, we consider two different approaches. In the former, a descriptor is computed from the body skeleton information provided by the Microsoft Kinect SDK [24] and fed to a pre-trained classifier. In the latter, we compare people point clouds by means of the fitness score obtained after an Iterative Closest Point (ICP) [5] registration. For tackling the problem of different poses people can have, we exploit the skeleton information for transforming a person point cloud to a standard pose before applying ICP.

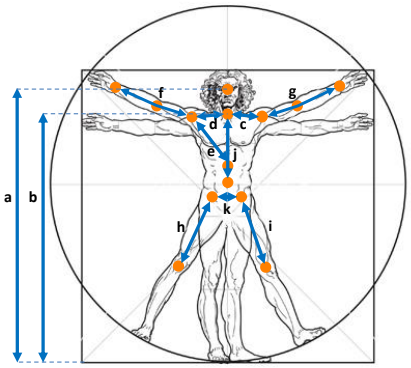
4.1 Feature-Based Re-Identification

In this section, our feature-based approach to person re-identification is described. In a first phase, as a subject is detected in front of the depth sensing device, the descriptor is extracted from the input channels. Our feature extraction step relies on the body skeleton obtained through the Kinect SDK, since the data is already available and computation is optimized.

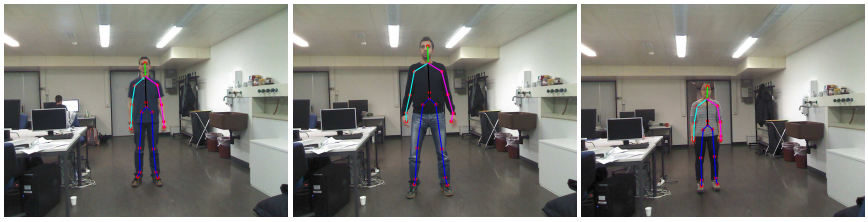
4.1.1 Skeleton Descriptor

The extraction of skeleton-based information is substantially the computation of a few limb lengths and ratios, using the 3D location of the body joints provided by the skeletal tracker. We extended the set of skeleton features used in [2], in order to collect measurements from all the human body. In particular, we extract the following 13 distances:

a) head height,
b) neck height,
c) neck to left shoulder distance,
d) neck to right shoulder distance,
e) torso to right shoulder distance,
f) right arm length,
g) left arm length,
h) right upper leg length,
i) left upper leg length,
j) torso length,
k) right hip to left hip distance,
l) ratio between torso length and right upper leg length (j/h),
m) ratio between torso length and left upper leg length (j/i).



All these distances are concatenated into a single skeleton descriptor $\mathbf{x}_S$. In Fig. 1, the skeleton computed with Microsoft Kinect SDK is reported for three very different people of our dataset, while in Fig. 2 and 3, we show how the value of some skeleton features varies along time when these people are still and walking, respectively. We also report the average standard deviation of these features for the people of the two testing sets. As expected, the heights of the head and the neck from the ground are the most discriminative features. What is more interesting is that the standard deviation of these features doubles for the walking test set with respect to the test set where people are still, thus suggesting that the skeleton joint positions are better estimated when people are static and frontal.



(a) (b) (c)

Fig. 1 Examples of estimated skeletons for three people of the testing videos of the *BIWI RGBD-ID* dataset.

When a person is seen from the side or from the back, Microsoft's skeletal tracking algorithm [24] does not provide correct estimates because it is based on a random forest classifier which has been trained with examples of frontal people only. For this reason, in this work, we discard frames with at least one not tracked joint[2]. Then, we keep only those where a face is detected [27] in the proximity of the

[2] Microsoft's SDK provides a flag for every joint stating if it is tracked, inferred or not tracked.

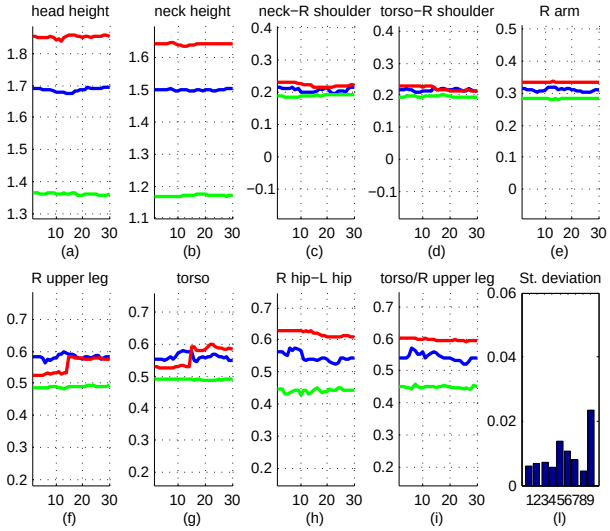


Fig. 2 (a-i) Estimated skeleton features for some frames of the *Still* test sequence for the three subjects of Fig. 1. Those subjects are represented by blue, red and green curves, respectively. In (l), the standard deviation of these features is reported.

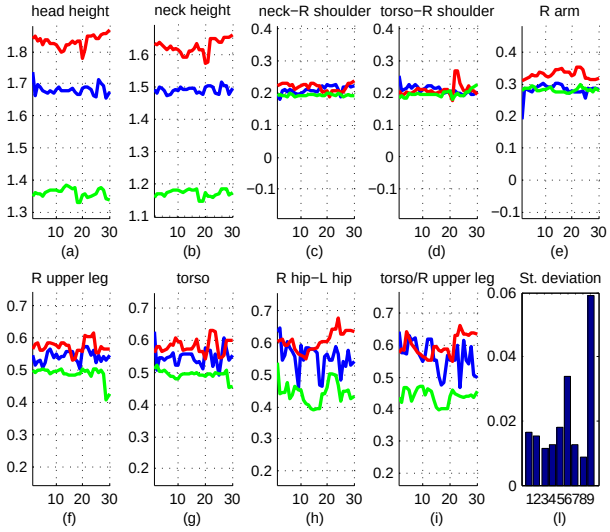


Fig. 3 (a-i) Estimated skeleton features for some frames of the *Walking* test sequence for the three subjects of Fig. 1. Those subjects are represented by blue, red and green curves, respectively. In (l), the standard deviation of these features is reported.

head joint position. This kind of selection is needed for discarding also those frames where the person is seen from the back, which come with a wrong skeleton estimation.

4.1.2 Classification

For classifying the descriptor presented in the previous section, we tested four different classification approaches. The first method compares descriptors extracted from the testing dataset with those of the training dataset by means of a Nearest Neighbor classifier based on the Euclidean distance. The second one consists in learning the parameters of a Support Vector Machine (SVM) [10] for every subject of the training dataset. As SVMs are originarily designed for binary classification, these classifiers are trained in a *One-vs-All* fashion: For a certain subject i, the descriptors computed on that subject are considered as positive samples while the descriptors computed on all the subjects except i are considered as negative samples.

The One-vs-All approach requires all the training procedure to be performed again if a new person is inserted in the database. This need makes the approach not suitable for a scenario where new people are inserted online for a subsequent re-identification. For this purpose, we also trained a *Generic SVM* which does not learn how to distinguish a specific person from all the others, but it learns how to understand if two descriptors have been extracted from the same person or not. The positive training examples which are fed to this SVM are of the form

$$pos = \left| d_1^i - d_2^i \right|, \tag{1}$$

where d_1^i and d_2^i are descriptors extracted from two frames containing the same subject i, while the negative examples are of the form

$$neg = \left| d_1^i - d_2^j \right|, \tag{2}$$

where d_1^i and d_2^j are descriptors extracted from frames containing different subjects. At testing time, the current descriptor d_{test} is compared to the training descriptors d_k^i of every subject i by using this Generic SVM for classifying the vector $\left| d_{test} - d_k^i \right|$ and the test descriptor is associated to the class for which the maximum SVM confidence is obtained.

Finally, we tested also a Naive Bayes approach: as a training stage, we computed mean and standard deviation of a normal distribution for every descriptor feature and for every person of the training dataset; at testing time, we used these data to calculate the likelihood with which a new descriptor could belong to each person in the training set.

4.2 Point Cloud Matching

The skeleton descriptor explained in Sec. 4.1.1 provides information about the characteristic lengths of the human body. However, it does not take into account many shape traits which are important for discriminating people with similar body lengths. In this section, we propose a process which takes the whole point cloud shape into account for the re-identification task. In particular, given two persons' point clouds, we try to align them and then compute a similarity score between the two. As a fitness score, we compute the average distance of the points of a cloud to the nearest points of the other cloud. If $\mathscr{P}_1$ and $\mathscr{P}_2$ are two point clouds, the fitness score of $\mathscr{P}_2$ with respect to $\mathscr{P}_1$ is then

$$f_{2\to 1} = \sum_{p_i \in \mathscr{P}_2} \|p_i - q_i^*\|, \tag{3}$$

where q_i^* is defined as

$$q_i^* = \arg\min_{q_j \in \mathscr{P}_1} \left\|p_i - q_j\right\|. \tag{4}$$

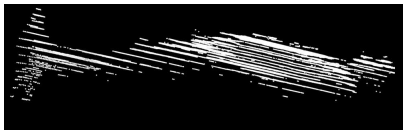
It is worth to notice that this fitness score is not symmetric, that is $f_{2\to 1} \neq f_{1\to 2}$.

For what concerns the alignment, the position and orientation of a reference skeleton joint, e.g. the hip center, is used to perform a rough alignment between the clouds to compare. Then, that alignment is refined by means of an ICP-based registration, which should converge in few iterations if the initial alignment is good enough. When the input point clouds have been aligned with this process, the fitness score between them should be minimum, ideally zero if they coincide or if $\mathscr{P}_2$ is contained in $\mathscr{P}_1$.

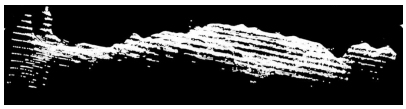
For the purpose of re-identification, this procedure can be used to compare a testing point cloud with the point clouds of the persons in the training set and to select the subject whose point cloud has the minimum fitness score when matched with the testing cloud. However, for this approach to work well, a number of problems should be taken into account, such as the quality of the depth estimates and the different poses people can assume.

4.2.1 Point Cloud Smoothing

3D point clouds acquired with consumer depth sensors have good resolution but the depth quantization step increasing quadratically with the distance does not allow to obtain smooth people point clouds beyond two meters from the sensor. In Fig. 4(a), the point cloud of a person three meters from the sensor is reported. It can be noticed that the point cloud results divided into slices produced by the quantization steps. As a pre-processing step, we improve the person point cloud by applying a voxel grid filter and a Moving Least Squares surface reconstruction method to obtain a smoothing, as reported in Fig. 4(b).



(a) Raw



(b) Smoothed

Fig. 4 (a) Raw person pointcloud at 3 meters of distance from the Kinect and (b) point cloud after the pre-processing step.

4.2.2 Point Cloud Transformation to Standard Pose

The point cloud matching technique we described is derived from the 3D object recognition research, where objects are supposed to undergo rigid transformations only. However, when dealing with moving people, the rigidity assumption does not hold any more, because people are articulated and they can appear in a very large number of different poses, thus these approaches would be doomed to fail.

Bronstein *et al.* ([7], [8]) tackle this problem by applying an isometric embedding which allows to get rid of pose variability (extrinsic geometry) by warping shapes to a canonical form where geodesic distances are replaced by Euclidean ones. In this space, an ICP matching is applied to estimate similarity between shapes. However, a geodesic masking which retains the same portion of every shape is needed for this method to work well. In particular, for matching people's shape, a complete and accurate 3D scan has to be used, thus partial views cannot be matched with a full model because they could lead to very different embeddings. Moreover, this approach needs to solve a complicated optimization problem, thus requiring several seconds to complete.

For these reasons, we studied a new technique which exploits the information provided by the skeleton for efficiently transforming people point clouds to a standard pose before applying the matching procedure. This result is obtained by roto-translating each body part according to the positions and orientations of the skeleton joints and links given by Microsoft's skeletal tracking algorithm.

A preliminary operation consists in segmenting the person's point cloud into body parts. Even if Microsoft's skeletal tracker estimates this segmentation as a first step and then derives the joints position, it does not provide to the user the result of the depth map labeling into body parts. For this reason, we implemented the reverse procedure for obtaining the segmentation of a person point cloud into parts by starting from the 3D positions of the body joints. In particular, we assign every point cloud point to the nearest body link. For a better segmentation of the torso and the arms, we added two further fictitious links between the hips and the shoulders.

Once we performed the body segmentation , we warp the pose assumed by the person to a new pose, which is called *standard pose*. The standard pose makes the point clouds of all the subjects directly comparable, by imposing the same orientation between the links. On the other hand, the joints/links position is person-dependent and is estimated from a valid frame of the person and then kept fixed. This approach allows the standard pose skeleton to adapt to the different body lengths of the subjects. This transformation consists in rototranslating the points belonging to every body part according to the corresponding skeleton link position and orientation[3]. In particular, every body part is rotated according to the corresponding link orientation and translated according to its joints coordinates. If Q_c is the quaternion representing the orientation of a link in the current frame given by the skeleton tracker and Q_s is the one expressing its orientation in standard pose, the whole rotation to apply can be computed as

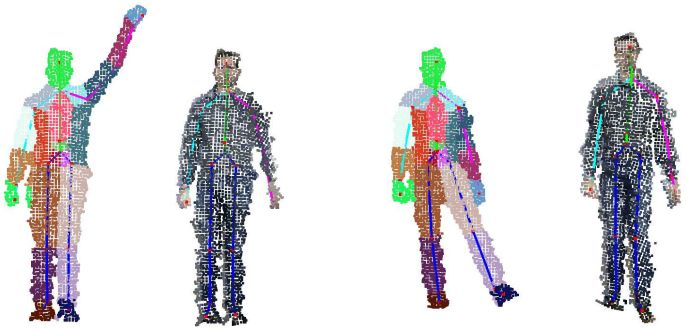
$$R = Q_s \left(Q_c\right)^{-1}, \tag{5}$$

while the full transformation applied to a point p can be synthesized as

$$p' = T_{V_s}\left(R\left(T_{V_c}\left(p\right)\right)\right), \tag{6}$$

where T_{V_c} and T_{V_s} are the translation vectors of the corresponding skeleton joint at the current frame and in the standard pose, respectively.

As the standard pose, we chose a typical frontal pose of a person at rest. In Fig. 5, we report two examples of person's point clouds before and after the transformation to standard pose. For the point cloud before the transformation, the body segmentation is shown with colors, while the points with RGB texture are reported for the transformed point cloud.



(a) (b)

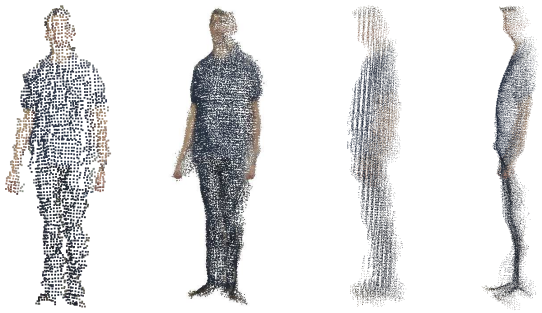
Fig. 5 Two examples of standard pose transformation. On the left, the body segmentation is shown with colors, on the right, the RGB texture is applied to the transformed point cloud.

[3] It is worth noting that all the links belonging to the torso have the same orientation, as the hip center.

It is worth noting that the process of rotating each body part according to the skeleton estimation can have two negative effects on the point cloud: some body parts can intersect each other and some gaps can appear around the joint centers. However, the parts intersection is tackled by voxel grid filtering the transformed point cloud, while the missing points do not represent a problem for the matching phase, since a test point cloud is considered to perfectly match a training point cloud if it is fully contained in it, as explained in Section 4.2.

4.2.3 Creation of Point Cloud Models

The transformation to standard pose is not only useful because it allows to compare people clouds disregarding their initial pose, but also because more point clouds belonging to the same moving person can be easily merged to compose a wider person model. In Fig. 6, a single person's point cloud (a) is compared with the model we obtained by merging together some point clouds acquired from different points of view and transformed to standard pose. It can be noticed how the union cloud is denser and more complete with respect to the single one. We also show, in Fig. 6(c) and (d), a side view of the person model when no smoothing is performed and when the smoothing of Sec. 4.2.1 is applied. Our approach is not focused on obtaining realistic 3D models for computer graphics, but on creating 3D models which can be useful for the re-identification task. In fact, these models can be used as a reference for matching new test point clouds with the people database. In particular, a point cloud model is created for every person from a sequence of frames where the person is turning around. Then, a new testing cloud can be transformed to standard pose and compared with all the persons' models by means of the approach described in Sec. 4.2. Given that, with Microsoft's skeletal tracker, we do not obtain valid frames if the person is seen from the back, we can only obtain 180° people models.



(a) (b) (c) (d)

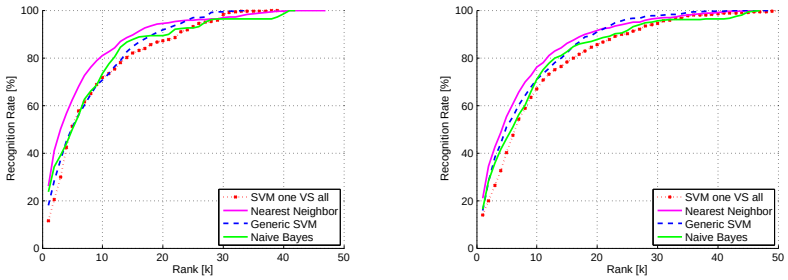
Fig. 6 (a) A single person's point cloud and (b) the point cloud model obtained by merging together several point clouds transformed to standard pose. Person's point cloud model (c) before and (d) after the smoothing described in Sec. 4.2.1.

5 Experiments

In this section, we report the experiments we carried out with the techniques described in Sec. 4. For evaluation purposes, we compute *Cumulative Matching Curves* (CMC) [13], which are commonly used for evaluating re-identification algorithms. For every k from 1 to the number of training subjects, these curves express the mean person recognition rate computed when considering a classification to be correct if the ground truth person appears among the subjects who obtained the k best classification scores. The typical evaluation parameters for these curves are the *rank-1* recognition rate and the *normalized Area Under Curve* (nAUC), which is the integral of the CMC. In this work, the recognition rates are separately computed for every subject and then averaged to obtain the final recognition rate.

5.1 Tests on the BIWI RGBD-ID Dataset

We present here some tests we performed on the *BIWI RGBD-ID* dataset. For the feature-based re-identification approach of Sec. 4.1, we extracted frame descriptors and trained the classifiers on the 50 sequences of the training set and we used them to classify the *Still* and *Walking* sequences of the 28 people of the testing set. In Fig. 7, we report the CMCs obtained on the *Still* and *Walking* testing sets when classifying the skeleton descriptor with the four classifiers described in Sec. 4.1.2. The best classifier for this kind of descriptor proved to be the Nearest Neighbor, which obtained a rank-1 recognition rate of 26.6% and a nAUC of 89.7% for the testing set where people are still and 21.1% and 86.6% respectively for the testing set with walking people.

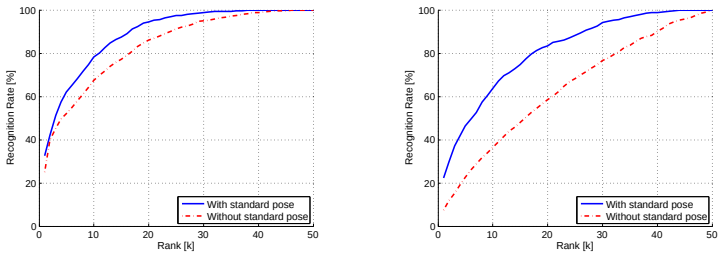


(a) Still (b) Walking

Fig. 7 Cumulative Matching Cuves obtained with the skeleton descriptor and different types of classifiers for the both the *Still* (a) and *Walking* (b) testing sets of the *BIWI RGBD-ID* dataset.

For testing the point cloud matching approach of Sec. 4.2, we built one point cloud model for every person of the training set by merging together point clouds

extracted from their training sequences and transformed to standard pose. At every frame, a new cloud is added and a voxel grid filter is applied to the union result for re-sampling the cloud and limiting the number of points. At the end, we exploit a moving least squares surface reconstruction method for obtaining a smoothing. At testing time, every person's cloud is transformed to standard pose, aligned and compared to the 50 persons' training models and classified according to the minimum fitness score $f_{test \rightarrow model}$ obtained. It is worth to notice that the fitness score reported in Eq. 3 correctly returns the minimum score (zero) if the test point cloud is contained in the model point cloud, while it would return a different score if the test cloud would only partially overlap the model. Also for this reason we chose to build the persons' models described above, i.e. by having training models covering 180° while the test point clouds are smaller and for this reason only cover portions of the training point clouds. In Fig. 8, we compare the described method with a similar matching method which does not exploit the point cloud transformation to standard pose. For the testing set with still people, the differences are small because people are often in the same pose, while, for the walking test set, the transformation to standard pose outperforms the method which does not exploit it, reaching a rank-1 performance of 22.4% against 7.4% and a nAUC of 81.6% against 64.3%.

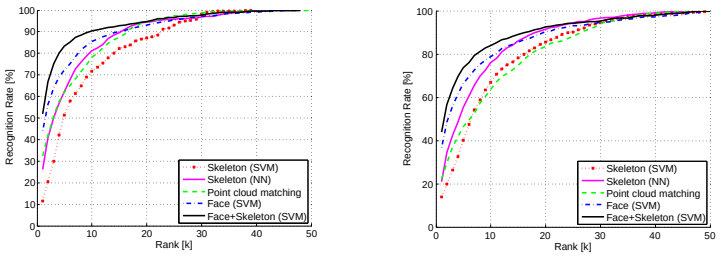


(a) Still (b) Walking

Fig. 8 Cumulative Matching Curves obtained with the point cloud matching approach with and without transformation to standard pose on the testing sets of the *BIWI RGBD-ID* dataset.

We compare the main approaches we described in Fig. 9. As a reference, we report also the results obtained with a face recognition technique. This technique extracts the subject's face from the RGB input using a standard face detection algorithm [27]. To increase the computational speed and decrease the number of false positives, the search region is limited to a small neighborhood of the 2D location of the head, as provided by the skeletal tracker. Once the face has been detected, a real-time method to extract the 2D location of 10 fiducials points is applied [11]. Finally, SURF descriptors [3] are computed at the location of the fiducials and concatenated forming a single vector. Unlike the skeleton descriptor, the face descriptor provided the best results with the One-VS-All SVM classifier, reaching 44% of rank-1 for the *Still* testing set and 36.7% for the *Walking* set. An advantage of the SVM classification is that descriptors referring to different features can be easily fused by con-

catenating them and leaving to the classifier the task to learn the suitable weights. We report, as an example, the results obtained with the concatenation of the face and skeleton descriptors which are then classified with the One-VS-All SVM approach. This method allows to further gain 8% of rank-1 for the *Still* test set and 7.2% for the *Walking* test set. In Table 1, all the numerical results are reported, together with those obtained by executing a three-fold cross validation on the training videos where two folds were used for training and one for testing. In the remaining experiments, all the training videos were used for training and all the testing data were used for testing. The point cloud matching technique performs slightly better than the skeleton descriptor classification for the *Still* test set and slightly worse for the *Walking* test set, thus proving to be useful too for the re-identification task.



(a) Still (b) Walking

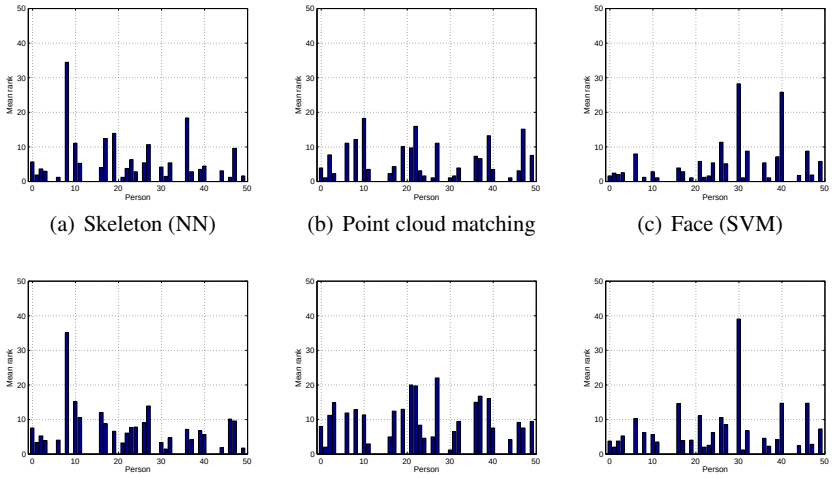
Fig. 9 Cumulative Matching Curves obtained with the main approaches described in this paper for the *BIWI RGBD-ID* dataset.

Table 1 Evaluation results obtained in cross validation and with the testing sets of the *BIWI RGBD-ID* dataset. The One-VS-All classifiers do not perform very well because the positive and negative samples are likely not well separated in feature space, due to the negative class being very widely spread. Although it is possible that pairwise classifiers may perform better, this would lead to a very large number of classifiers, which may be impractical given the number of classes. This non-separability at the category level is supported by the good performance of the nearest neighbor classifier, which further suggests that there are overlaps among categories, but locally some classification is possible.

	Cross validation		**Test - Still**		**Test - Walking**	
	Rank-1	**nAUC**	**Rank-1**	**nAUC**	**Rank-1**	**nAUC**
Skeleton (SVM)	47.5%	96.1%	11.6%	84.5%	13.8%	81.7%
Skeleton (NN)	80.5%	98.2%	26.6%	89.7%	21.1%	86.6%
Point cloud matching	93.7%	99.6%	32.5%	89.0%	22.4%	81.6%
Face (SVM)	97.8%	99.4%	44.0%	91.0%	36.7%	87.6%
Face+Skeleton (SVM)	98.4%	99.5%	52.0%	93.7%	43.9%	90.2%

For analyzing how the re-identification performance changes for the different people of our dataset, we report in Fig. 10 the histograms of the mean ranking for

every person of the testing dataset, which is the average ranking at which the correct person is classified. The missing values in the x axis are due to the fact that not all the training subjects are present in the testing set. It can be noticed that there is a correspondence between the mean ranking obtained in the *Still* testing set and that obtained in the *Walking* test set. On the contrary, it is also clear that different approaches lead to mistakes on different people, thus showing to be partially complementary.



(a) Skeleton (NN) (b) Point cloud matching (c) Face (SVM)

(d) Skeleton(NN) (e) Point cloud matching (f) Face (SVM)

Fig. 10 Mean ranking histograms obtained with different techniques for every person of the *Still* (top row) and *Walking* (bottom row) test sets of the *BIWI RGBD-ID* dataset.

5.2 Tests on the RGB-D Person Re-Identification Dataset

As explained in Section 3, the *RGB-D Person Re-Identification* dataset is the only other public dataset for person re-identification using RGB-D data. Unfortunately, there are only few examples available for each of the subjects, which makes the use of many machine learning techniques, including SVMs trained with a One-VS-All approach, quite complicated. However, given that the Generic SVM described in Sec. 4.1.2 is one for all the subjects, we had enough examples to train it correctly. In Table 2, we compare the results reported in [2] with our results obtained when classifying the skeleton descriptor with the Nearest Neighbor and the Generic SVM. Unfortunately, the authors of [2] report performances only in terms of normalized Area Under Curve (nAUC) of the Cumulative Matching Curve (CMC), thus their rank-1 scores are not available except for one result that can be inferred from a fig-

ure. The classification of our skeleton descriptor with the Generic SVM performed better than [2] and of our Nearest Neighbor classifier for the tests which do not involve the *Collaborative* set, where people walk with open arms. We also tested the geodesic features the authors propose, but they did not provide substantial improvement to the skeleton alone. We did not test the point cloud matching and the face recognition techniques on this dataset because the links orientation information was not provided and the face in the RGB image was blurred.

Table 2 Evaluation results on the *RGB-D Person Re-Identification* dataset.

Training	Testing	[2]		Ours - NN		Ours - Generic SVM	
		Rank-1	**nAUC**	**Rank-1**	**nAUC**	**Rank-1**	**nAUC**
Collaborative	**Walking1**	N/A	90.1%	7.8%	81.1%	5.3%	79.0%
Collaborative	**Walking2**	13%	88.9%	4.8%	81.3%	4.1%	78.6%
Collaborative	**Backwards**	N/A	85.6%	4.6%	78.8%	3.6%	76.0%
Walking1	**Walking2**	N/A	91.8%	28.6%	89.9%	35.7%	92.8%
Walking1	**Backwards**	N/A	88.7%	17.8%	82.7%	18.5%	90.6%
Walking2	**Backwards**	N/A	87.7%	13.2%	84.1%	22.3%	91.6%

5.3 Multi-Frame Results

The re-identification methods we described in this work are all based on a one-shot re-identification from a single test frame. However, when more frames of the same person are available, the results obtained for each frame can be merged to obtain a sequence-wise result. In Table 3, we compare on our dataset the single-frame rank-1 performances with what can be obtained with a simple multi-frame reasoning, that is by associating each test sequence to the subject voted by the highest number of frames. On average, this voting scheme allows to obtain a performance improvement of about 8-10%. The Nearest Neighbor classification of the skeleton descriptor for the *Walking* test set seems to benefit most from this approach, thus its rank-1 almost doubles. The best performance is again obtained with the SVM classification of the combined face and skeleton descriptors, which reaches 67.9% of rank-1 for both the testing sets.

5.4 Runtime Performance

The feature-based re-identification method of Sec. 4.1 exploits information which is already pre-computed by Microsoft Kinect SDK and classification methods which takes less than a millisecond to classify one frame, thus the runtime performance is only limited by the sensor frame rate and by the face detection algorithm used to select frames with a valid skeleton, which runs at more than 20fps with a C++

Table 3 Rank-1 results with the single-frame and the multi-frame evaluation for the testing sets of the *BIWI RGBD-ID* dataset.

	Cross validation		**Test - Still**		**Test - Walking**	
	Single	**Multi**	**Single**	**Multi**	**Single**	**Multi**
Skeleton (SVM)	47.5%	66.0%	11.6%	10.7%	13.8%	17.9%
Skeleton (NN)	80.5%	100%	26.6%	32.1%	21.1%	39.3%
Point cloud matching	93.7%	100%	32.5%	42.9%	22.4%	39.3%
Face (SVM)	97.8%	100%	44.0%	57.1%	36.7%	57.1%
Face+Skeleton (SVM)	98.4%	100%	52.0%	67.9%	43.9%	67.9%

implementation on a standard workstation with an Intel Core i5-3570k@3.40GHz processor.

In Table 4, the runtime of the single algorithms needed for the point cloud matching method of Sec. 4.2 are reported. The most demanding operation is the matching between the test point cloud transformed to standard pose and the models of every subject in the training set, which takes 250ms for performing 50 comparisons. The overall frame rate is then of about 2.8fps, which suggests that also this approach could be used in a real-time scenario with further optimization and with a limited number of people in the database.

Table 4 Runtime performance of the algorithms used for the point cloud matching method.

	time (ms)
Face detection	42.19
Body segmentation	3.03
Transformation to standard pose	0.41
Filtering and smoothing	56.35
ICP and fitness scores computation	254.34

6 Conclusions and Directions for Future Work

In this chapter, we have compared two different techniques for one-shot person re-identification with soft biometric cues obtained through a consumer depth sensor. The skeleton information is used to build a descriptor which can then be classified with standard machine learning techniques. Moreover, we also proposed to identify subjects by comparing their global body shape. For this purpose, we described how to warp point clouds to a standard pose in order to allow a rigid comparison based on a typical ICP fitness score. We also proposed to use this transformation for obtaining a 3D body model which can be used for re-identification from a series of point clouds of the subject while moving freely.

We tested the proposed algorithms on a publicly available dataset and on the newly created *BIWI RGBD-ID* dataset, which contains 50 training videos and 56 testing sequences with synchronized RGB, depth and skeleton data. Experimental results show that both the skeleton and the shape information can be used for effec-

tively re-identifying subjects in a non-collaborative scenario, because similar results have been obtained with these two approaches.

As future work, we envision to study techniques for combining skeleton classification and point cloud matching results into a common single re-identification framework.

7 Acknowledgements

The authors would like to thank all the people at the BIWI laboratory of ETH Zurich who took part in the *BIWI RGBD-ID* dataset.

References

1. Apostoloff, N., Zisserman, A.: Who Are You? - Real-time Person Identification. In: British Machine Vision Conference (2007)
2. Barbosa, B.I., Cristani, M., Del Bue, A., Bazzani, L., Murino, V.: Re-identification with rgb-d sensors. In: First International Workshop on Re-Identification (2012)
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer Vision and Image Understanding **110**(3), 346–359 (2008)
4. Bedagkar-Gala, A., Shah, S.: Multiple person re-identification using part based spatio-temporal color appearance model. In: Computational Methods for the Innovative Design of Electrical Devices'11, pp. 1721 –1728 (2011)
5. Besl, P.J., McKay, N.: A method for registration of 3-d shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence **14**, 239–256 (1992)
6. Bowyer, K.W., Chang, K., Flynn, P.: A survey of approaches and challenges in 3d and multi-modal 3d + 2d face recognition. Computer Vision and Image Understanding **101**(1), 1–15 (2006)
7. Bronstein, A.M., Bronstein, M.M., , Kimmel, R.: Three-dimensional face recognition. International Journal of Computer Vision **64**, 5–30 (2005)
8. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Topology-invariant similarity of nonrigid shapes. International Journal of Computer Vision **81**, 281–301 (2009)
9. Brunelli, R., Falavigna, D.: Person identification using multiple cues. IEEE Trans. on Pattern Analysis and Machine Intelligence **17**(10), 955 –966 (1995)
10. Cortes, C., Vapnik, V.N.: Support-vector networks. Machine Learning **20** (1995)
11. Dantone, M., Gall, J., Fanelli, G., Gool, L.V.: Real-time facial feature detection using conditional regression forests. In: IEEE Conf. on Computer Vision and Pattern Recognition (2012)
12. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360 –2367 (2010)
13. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision, vol. 5302, pp. 262–275 (2008)
14. Hong, L., Jain, A., Pankanti, S.: Can multibiometrics improve performance? In: Proc. IEEE Workshop on Automatic Identification Advanced Technologies, pp. 59–64 (1999)
15. Jain, A.K., Dass, S.C., Nandakumar, K.: Can soft biometric traits assist user recognition? Proc. SPIE, Biometric Technology for Human Identification **5404**, 561–572 (2004)
16. Lee, S.U., Cho, Y.S., Kee, S.C., Kim, S.R.: Real-time facial feature detection for person identification system. In: Machine Vision and Applications, pp. 148–151 (2000)
17. Leyvand, T., Meekhof, C., Wei, Y.C., Sun, J., Guo, B.: Kinect identity: Technology and experience. Computer **44**(4), 94 –96 (2011)

18. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: IEEE International Workshop on CVPR for Human Communicative Behavior Analysis (in conjunction with CVPR 2010), San Francisco, CA, (2010)
19. Ober, D., Neugebauer, S., Sallee, P.: Training and feature-reduction techniques for human identification using anthropometry. In: Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, pp. 1 –8 (2010)
20. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: A comprehensive multimodal human action database. In: Proc. of the IEEE Workshop on Applications on Computer Vision (2013)
21. Preis, J., Kessel, M., Werner, M., Linnhoff-Popien, C.: Gait recognition with kinect. In: Proceedings of the First Workshop on Kinect in Pervasive Computing (2012)
22. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters **24**, 2115–2125 (2003)
23. Satta, R., Pala, F., Fumera, G., Roli, F.: Real-time appearance-based person re-identification over multiple Kinect cameras. In: VisApp (2013)
24. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1297–1304 (2011)
25. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from rgbd images. In: International Conference on Robotics and Automation (2012)
26. Velardo, C., Dugelay, J.L.: Improving identification by pruning: A case study on face recognition and body soft biometric. In: International Workshop on Image and Audio Analysis for Multimedia Interactive Services, pp. 1–4 (2012)
27. Viola, P.A., Jones, M.J.: Robust real-time face detection. In: International Conference on Computer Vision, p. 747 (2001)
28. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Ma, Y.: Towards a practical face recognition system: Robust registration and illumination by sparse representation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 597 –604 (2009)
29. Wang, C., Zhang, J., Pu, J., Yuan, X., Wang, L.: Chrono-gait image: A novel temporal template for gait recognition. In: Proceedings of the 11th European Conference on Computer Vision, pp. 257–270 (2010)
30. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)
31. Wang, S., Lewandowski, M., Annesley, J., Orwell, J.: Re-identification of pedestrians with variable occlusion and scale. In: International Conference on Computer Vision Workshops, pp. 1876 –1882 (2011)
32. Wolf, C., Mille, J., Lombardi, E., Celiktutan, O., Jiu, M., Baccouche, M., Dellandrea, E., Bichot, C.E., Garcia, C., Sankur, B.: The liris human activities dataset and the icpr 2012 human activities recognition and localization competition. Tech. Rep. RR-LIRIS-2012-004 (2012)
33. Zhang, H., Parker, L.E.: 4-dimensional local spatio-temporal features for human activity recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2044 –2049 (2011)
34. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Comput. Surv. **35**(4), 399–458 (2003)
35. Zhu, P., Zhang, L., Hu, Q., Shiu, S.: Multi-scale patch based collaborative representation for face recognition with margin distribution optimization. In: European Conference on Computer Vision, pp. 822–835 (2012)

Index

BIWI RGBD-ID Dataset, 1, 4, 13
Body segmentation, 11

Cumulative Matching Curves, 13

Depth images, 5

Face recognition, 3, 14
Feature-based, 2, 5, 13
Fitness score, 5, 9

Kinect, 2–4, 6

Moving Least Squares, 9, 14
Multi-frame, 17

Naive Bayes, 8
Nearest Neighbor classifier, 8, 17
Normalized Area Under Curve, 13

One-shot re-identification, 2

Person model, 12
Point cloud, 1, 2, 5, 9, 10
Point cloud matching, 10, 13, 18

Rank-1, 13, 14, 16, 17
Real-time, 3, 18
RGB-D Person Re-Identification dataset, 16

Shape, 1, 3, 9
Skeletal tracker, 5, 10, 12, 14
Soft biometrics, 2
Standard pose, 1, 2, 5, 10, 14, 18
Support Vector Machine, 8

Union cloud, 12