3D Reconstruction of freely moving persons for re-identification with a depth sensor

Conference Paper *in* Proceedings - IEEE International Conference on Robotics and Automation · May 2014

DOI: 10.1109/ICRA.2014.6907518

4 authors, including:

Matteo Munaro
University of Padova
41 PUBLICATIONS **357** CITATIONS
SEE PROFILE

Andrea Fossati
ETH Zurich
19 PUBLICATIONS **473** CITATIONS
SEE PROFILE

Luc Van Gool
ETH Zurich
1,022 PUBLICATIONS **60,859** CITATIONS
SEE PROFILE

3D Reconstruction of Freely Moving Persons for Re-Identification with a Depth Sensor

Matteo Munaro[1], Alberto Basso[1], Andrea Fossati[2], Luc Van Gool[2], Emanuele Menegatti[1]

***Abstract*— In this work, we describe a novel method for creating 3D models of persons freely moving in front of a consumer depth sensor and we show how they can be used for long-term person re-identification. For overcoming the problem of the different poses a person can assume, we exploit the information provided by skeletal tracking algorithms for warping every point cloud frame to a standard pose in real time. Then, the warped point clouds are merged together to compose the model. Re-identification is performed by matching body shapes in terms of whole point clouds warped to a standard pose with the described method. We compare this technique with a classification method based on a descriptor of skeleton features and with a mixed approach which exploits both skeleton and shape features. We report experiments on two datasets we acquired for RGB-D re-identification which use different skeletal tracking algorithms and which are made publicly available to foster research in this new research branch.**

I. Introduction

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

Global shape is a *soft biometric* feature [1] that can allow to discriminate between people in a long-term time span when other cues are not available or are not enough for obtaining accurate results. In recent years, it became easily available with consumer RGB-D sensors, such as Microsoft Kinect, which equip the majority of modern mobile robots. However, little effort has been spent so far to develop methods for people re-identification based on this particular cue. In order to identify a person within a training set of known people given a partial point cloud obtained by a depth sensor, complete 3D models of the people in the training set are necessary. Moreover, the matching between test clouds and training models can fail because a person is not rigid and can assume a great variety of poses. The main sources of shape variability among clouds belonging to the same person are differences in pose and clothing between training and testing point clouds. If we assume that the differences in clothing shape are negligible, we could compare people as we do for rigid objects if they all had the same pose. The main idea investigated with this work is then to exploit the information provided by state of the art skeletal tracking algorithms to transform persons point clouds to a neutral pose. The proposed method is useful both for creating full body models which can be used as training set and for matching new point clouds with the training models. It is very efficient, so that both training and testing can be done online onboard a mobile robot. Moreover, our approach does not require the cooperation of the scanned person, that can perform arbitrary motions.

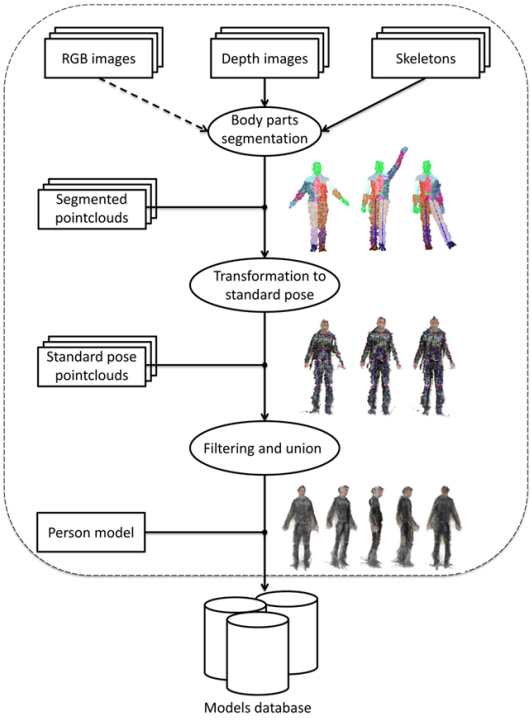


Fig. 1. Illustration of the pipeline we developed for creating full 3D models of freely moving persons. RGB information is added to the point cloud models only for a better visualization, but in this work it is not used for matching.

The contribution of this paper is three-fold: On one hand we propose a novel technique for exploiting skeleton information to efficiently transform persons' point clouds to a standard pose. On the other hand, we describe how to use these transformed point clouds for composing 3D models of freely moving people which can be used for re-identification, by means of an Iterative Closest Point (ICP) matching with new test clouds. Finally, we propose a method to combine the ICP matching scores with a descriptor of body skeleton lengths which improves the recognition rates obtainable by the stand-alone approaches. We compare these

[1]Matteo Munaro, Alberto Basso and Emanuele Menegatti are with the Intelligent Autonomous Systems Laboratory (IAS-Lab) at the University of Padua, Padua, IT. {munaro, bassoal1, emg}@dei.unipd.it.

[2]Andrea Fossati and Luc Van Gool are with the Computer Vision Laboratory at ETH Zurich, Zurich, CH. {fossati, vangool}@vision.ee.ethz.ch.

techniques with skeleton-based and face-based baselines on the newly acquired *BIWI RGBD-ID* and *IAS-Lab RGBD-ID* re-identification datasets, which exploit two different state of the art skeletal tracking algorithms.

II. Related Work

A. Depth-based and multi-modal re-identification

Due to the very recent availability of cheap depth sensing devices, only a few works exist that focus on identification using such multi-modal input. Most of these works rely on combinations of soft biometric features, that are in general not discriminative enough to identify a subject, but can be very powerful if combined with other traits. In [2], 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 [3] 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 [4], 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 technique purely relies on appearance-based features. The authors of [5] 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 a few extensive 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. The dataset used in the paper has also been made publicly available, but this does not contain facial information of the subjects and skeleton links orientation, in contrast with the datasets proposed within this paper. Also Kinect Identity [6], 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 small. A recent work also applied gait recognition [7] techniques to classify sequences of descriptors of joints information obtained with Kinect skeletal tracker.

Bronstein *et al.* ([8], [9]) exploit global body shape for re-identification and tackle the person matching 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. Our method, instead, exploits the information provided by a skeletal tracking algorithm for rapidly transforming persons point clouds to a standard pose, so that techniques studied for reconstructing objects can then be applied.

B. 3D Reconstruction of people

To recognize people based on their global shape, full 3D models have to be composed from sequences of depth frames, so that they can be used as training set. Even though human body reconstruction from multiple static sensors has been thoroughly studied, little literature exists on person reconstruction from a single moving sensor. As discussed in Sec. I, people are articulated and locally deformable, thus recent techniques for real-time RGB-D reconstruction [10], [11], which assume rigid scenes, are doomed to fail if the person is moving. In [12], the authors propose a method which exploits the SCAPE body model for obtaining 3D models of people from range scans acquired by a Microsoft Kinect. However, their approach is computationally expensive and still requires the person to be collaborative during the scanning since it is targeted to fitness and apparel applications.

III. RGB-D Re-Identification Datasets

The vast majority of publicly available RGB-D datasets are targeted to human activity analysis and action recognition, and for this reason they are generally composed by many gestures performed by few subjects [13], [14], [15], [16], [17], [18]. The only dataset explicitly thought for the RGB-D re-identification task has been proposed in [5]. It consists of 79 different subjects collected in 4 different scenarios. However, this dataset contains very few frames for each subject, faces are blurred for privacy reasons and skeleton links orientation are not available. To overcome these limitations, we collected our own RGB-D re-identification datasets.

A. BIWI RGBD-ID Dataset

The *BIWI RGBD-ID Dataset*[1] consists of video sequences of 50 different subjects, performing a certain routine of motions and walks in front of a Kinect. The dataset includes synchronized RGB images (at 1280 $\times$ 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 8-10fps and last about one minute for every subject.

Moreover, we have collected a *Still* and a *Walking* test sequence for 28 subjects already present in the dataset. In the *Walking* video, every person performs two walks frontally and other two walks diagonally with respect to the Kinect. These have been collected on a different day and therefore

[1] `http://robotics.dei.unipd.it/reid`.

most subjects are dressed differently. These sequences are also shot in a different location than the studio room where the training dataset had been collected.

B. IAS-Lab RGBD-ID Dataset

The skeletal tracking algorithm provided by Microsoft SDK gives the best accuracy, but does not allow to estimate the skeleton of non-frontal people. For this reason, we also collected 33 sequences[2] of 11 people with the OpenNI SDK[3] and the NITE[4] middleware, which does not have this limitation. For every subject, the first (*Training*) and the second (*TestingA*) sequences were acquired with people wearing different clothes, while the third one (*TestingB*) was collected in a different room, but with the same clothes as in the first sequence.

IV. Point Cloud Matching (PCM)

In this section, we propose a method which takes into account the whole human body point cloud for the re-identification task. In particular, given two 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 $\mathcal{P}_1$ and $\mathcal{P}_2$ are two point clouds, the fitness score of $\mathcal{P}_2$ with respect to $\mathcal{P}_1$ is then

$$f_{2\to1} = \frac{1}{N} \sum_{p_i \in \mathcal{P}_2} \|p_i - q_i^*\|, \tag{1}$$

where q_i^* is defined as

$$q_i^* = \arg\min_{q_j \in \mathcal{P}_1} \|p_i - q_j\| . \tag{2}$$

and N is the number of points in $\mathcal{P}_2$. It is worth noting that this fitness score is not symmetric, i.e. $f_{2\to1} \neq f_{1\to2}$.

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 point clouds to compare. Then, the alignment is refined by means of an ICP-based registration step [19], which converges in few iterations if the initial alignment is good enough. When the input point clouds have been aligned through this process, the fitness score between them should be minimum, ideally zero if they coincide or if $\mathcal{P}_2$ is contained in $\mathcal{P}_1$.

For the purpose of re-identification, this procedure is used to compare a testing point cloud with the point cloud models obtained from the training set as we will describe in Sec. VI, and to select the training subject whose point cloud model has the minimum fitness score when matched with the testing cloud.

[2] http://robotics.dei.unipd.it/reid.
[3] http://www.openni.org/openni-sdk.
[4] http://www.openni.org/files/nite.

V. Point Cloud Transformation to Standard Pose

Composing 3D models of persons usually requires multiple cameras or at least that the person remains still during the scanning process, as typical reconstruction techniques for rigid objects are usually exploited [20]. 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 making these approaches not suitable.

To overcome this problem, we exploit the information provided by a skeletal tracking algorithm to efficiently warp persons point clouds to a standard pose, so that techniques studied for reconstructing rigid objects can then be applied.

This result is obtained by rototranslating each body part according to the positions and orientations of the skeleton joints and links given by the skeletal tracking algorithm. A preliminary operation consists in segmenting the person's point cloud into body parts, so that a different transformation can be applied to each body part. Even if Microsoft's skeletal tracker estimates this segmentation as a first step and then derives the joints position, its output is not provided to the user. For this reason, we implemented the reverse procedure to obtain the segmentation of a 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 among those provided by the skeletal tracking algorithm. For a better segmentation of the torso and the arms, we added two further fictitious links between the hips and the shoulders. The body links considered and an example of body segmentation are depicted in Fig. 2(a).

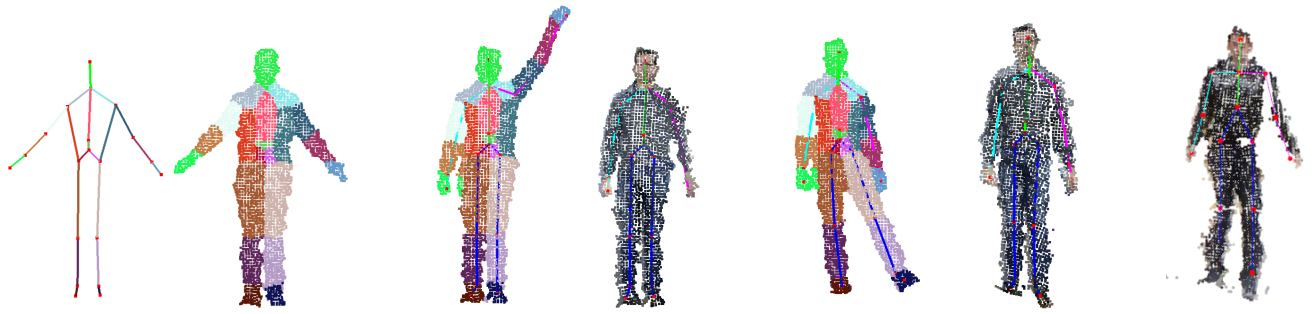
After the body segmentation step, the pose assumed by the subject is warped to a new pose, which is called *standard pose*. The standard pose makes the point clouds of all the subjects directly comparable, by imposing the same orientation between the links. On the other hand, each link length is person-dependent and is estimated from a valid frame of the person and then kept fixed. The transformation consists in rototranslating the points belonging to each body part according to the corresponding skeleton link position and orientation[5]. 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 containing the orientation of a link in the current frame given by the skeletal tracker, and Q_s is the one expressing its orientation in standard pose, the whole rotation to apply can be computed as

$$R = Q_s \left(Q_c\right)^{-1}, \tag{3}$$

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

$$p' = T_{V_s}\left(R\left(\left(T_{V_c}\right)^{-1}(p)\right)\right), \tag{4}$$

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



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

Fig. 2. (a) Body links considered and body segmentation obtained. (b-c) 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 point cloud obtained after the transformation. (d) Example of bad transformation to standard pose caused by a bad estimation of the skeleton links.

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. 2(b-c), we report two examples of point clouds before and after the transformation to standard pose.

It is worth noting that the point cloud transformation to standard pose, that is the process of rototranslating 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, that is by applying a smart downsampling which returns point clouds with constant density and reduces the effect of noise. On the other hand, 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 Sec. IV. In Fig. 2(d), we report an example of bad result of point cloud transformation to standard pose caused by the fact that the skeletal tracker estimated a wrong orientation for the links corresponding to the right leg and the left arm.

VI. Creation of Human Body 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 more complete body model. In Fig. 3(a), a single person point cloud is compared with the model we obtained by merging together several point clouds acquired while the person was moving and transformed to the standard pose. Moreover, we show the union cloud we obtained without any smoothing and after a voxel grid filter and a Moving Least Squares surface reconstruction method are applied. It can be noticed how the union cloud is denser and more complete with respect to the single cloud.

For the re-identification task, we create a point cloud model for every person from a sequence of training frames where the person is moving freely. For tackling the problem of noisy skeleton estimates, every body part is further registered to the model already composed by means of a local ICP algorithm. 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 (Fig. 3(b) and (c)), while with the OpenNI SDK we can reconstruct a $360°$ people model (Fig. 3(d) and (e)). It can be noticed how the front and back of the models obtained with the OpenNI SDK are too close to each other. This is due to a limitation of the skeletal tracking algorithm, which does not estimate correctly the joints position inside the bones. However, this offset could be measured and compensated as a future work. In the video attachment to this paper, the models we created for all the people of the *BIWI RGBD-ID Dataset* and the *IAS-Lab RGBD-ID Dataset* are shown.

VII. Skeleton descriptor and combination with Point Cloud Matching

In this section, we compare the proposed shape-based technique with a feature-based method, similar to [5], which computes a descriptor composed of skeleton links lengths and joints distances to the ground and classifies it by means of a Nearest Neighbor classifier based on the Euclidean distance.

In particular, our skeleton-based descriptor is a vector of 13 elements: the 11 distances shown in Fig. 4 and two length ratios, that are between the torso length and the right/left upper leg lengths (j/h and j/i).

Since the skeleton lengths and the body shape are complementary features, a combination of these approaches could lead to results superior to those obtained with the single techniques. For this reason, we also propose a mixed approach which uses the fitness scores obtained by matching the testing cloud with the training models as weights for the distances between the testing and training skeleton descriptors. Given f^i_{ICP} as the fitness score obtained by comparing a test point cloud warped to standard pose with the i^{th} model of the training dataset and be d^i_{skel} the minimum distance between the test skeleton descriptor and the descriptors of the i^{th} person of the training set. Then, according to our combined

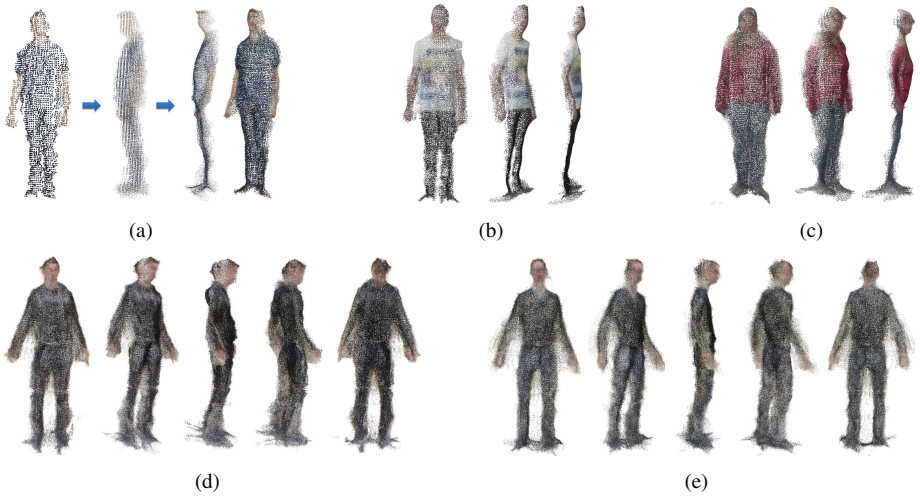


Fig. 3. (a) Steps of model creation (from left to right: a single person point cloud, the union of point clouds before and after smoothing); (b-c) examples of 180° persons models obtained with Kinect SDK; (d-e) examples of 360° persons models obtained with OpenNI SDK.

approach the distance of the test frame to the i^{th} person of the training set is computed as

$$d^i_{PCM+skel} = f^i_{ICP} \cdot d^i_{skel}, \quad i = 1...N_{train}, \qquad (5)$$

where N_{train} is the number of subjects in the training dataset. This procedure weights the skeleton classification differently for every person of the training set according to how well the current testing point cloud matches the training models obtained as described in Sec. VI. The current test frame is then associated to the training person obtaining the minimum $d^i_{PCM+skel}$. This weighting scheme allows to easily combine the influence of both techniques at the feature level. Other methods could be developed which combine them at the classification level.

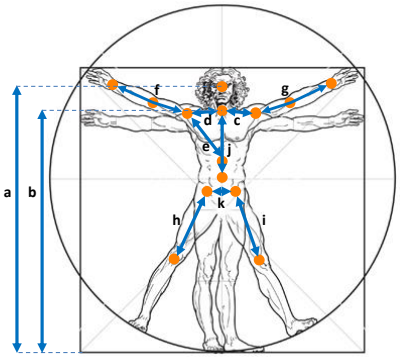


Fig. 4. Illustration of the links lengths and joints distances to the ground which constitutes the skeleton descriptor.

VIII. Experiments

We present some person re-identification experiments we have performed on the datasets described in Sec. III. For evaluation purposes, we compute *Cumulative Matching Characteristic (CMC) Curves* [21], 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 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.

A. Tests on the BIWI RGBD-ID Dataset

Microsoft's skeletal tracking algorithm is based on a random forest classifier which has been trained with examples of frontal people only, thus it does not provide correct estimates when the person is seen from the side or from the back. For this reason, in this work, we discarded the frames with at least one not tracked joint[6]. Then, we kept only those where a face was detected in the proximity of the head joint position. This kind of selection is needed to discard those frames where the person is seen from the back, which come with a wrong skeleton estimation.

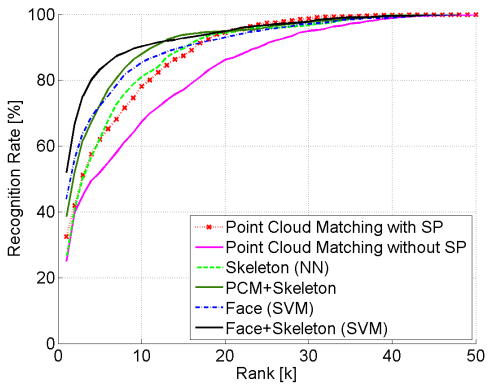
To test the point cloud matching approach of Sec. IV, we built a point cloud model for every person of the training set by merging together point clouds extracted from their training sequences and transformed to standard pose. At every frame, a new cloud was added and a voxel grid filter was applied to the union result for re-sampling the cloud and limiting the number of points. At the end, we exploited a moving least squares surface reconstruction method for smoothing. At testing time, every person cloud was transformed to standard pose, aligned and compared to the 50 persons training models and classified according to the minimum fitness score $f_{test \to model}$ obtained. It is worth noting that the fitness score reported in Eq. 1 correctly

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

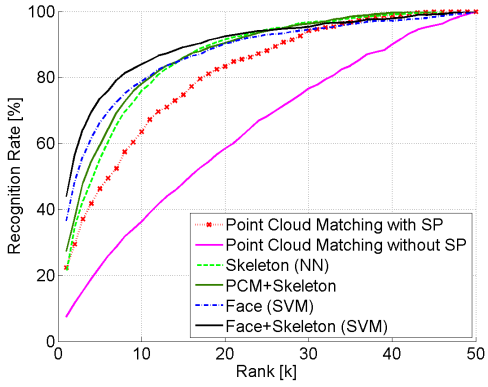
returns the minimum score (zero) if the test point cloud is contained in the model point cloud, while it would return a different score if the test cloud would only partially overlap the model. Also for this reason, we chose to build the persons models described in Sec. VI. In Fig. 5, we compare the 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 minor because people are often in the same pose, while, for the walking test set, the transformation to standard pose considerably 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%. We report also the results we obtained by classifying the skeleton descriptor and with the combined *PCM+Skeleton* approach of Sec. VII. It can be noticed how the point cloud matching technique obtained a slightly better rank-1 performance with respect to the skeleton classification, while the combined approach outperformed both methods, thus proving to exploit the complementarity of joint lengths and body shape. As a further 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 [22]. Once the face has been detected, a real-time method to extract the 2D location of 10 fiducials points is applied [23]. Finally, SURF descriptors [24] are computed at the location of the fiducials and concatenated forming a single vector. Unlike the skeleton descriptor, the face descriptor has been classified with a One-VS-All Support Vector Machine (SVM) classifier, reaching 44% of rank-1 for the *Still* testing set and 36.7% for the *Walking* set. Finally, by concatenating the skeleton and face descriptors and classifying them with a One-VS-All SVM, a further 8% gain of rank-1 for the *Still* test set and 7.2% for the *Walking* test set can be obtained. In Tab. I, all the numerical results are reported with also the cross validation outputs.

The re-identification methods we described 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 Tab. I, we also report the rank-1 performances which can be obtained with a simple multi-frame reasoning (*R1 Multi*), 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 10-20%. 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, while the combined PCM+skel approach obtained a rank-1 of 46.4%, thus proving to be the best option when face is not available.

To analyze how the re-identification performance differs for the different subjects of our dataset, we report in Fig. 6 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 the training set is composed of 50



(a) Still



(b) Walking

Fig. 5. Cumulative Matching Characteristic Curves obtained with the main approaches described in this paper for the *BIWI RGBD-ID Dataset*.

people, while the testing set contains 28 people out of these, thus we are showing re-identification results only for those people which 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. It is also clear that different approaches lead to mistakes on different people, thus showing to be partially complementary.

B. Tests on the IAS-Lab RGBD-ID Dataset

For the *IAS-Lab RGBD-ID Dataset*, we performed the same evaluation described for the *BIWI RGBD-ID Dataset*. However, we used also the frames where persons were seen from the back. The OpenNI SDK provides much noisier estimates than the Kinect SDK and sometimes estimates implausible poses, e.g. with legs turned backwards when the torso is turned forwards or with legs crossing and compenetrating. These erroneous frames are automatically detected by our algorithm by comparing the orientation of the torso with those of the legs and by comparing the relative orientation of the two legs against the physically possible range of values for a human being. Frames with implausible values are then discarded. In Fig. 7, we report the CMC curves obtained with the point cloud matching, the skeleton descriptor and the combined approach with the *IAS-Lab*

TABLE I

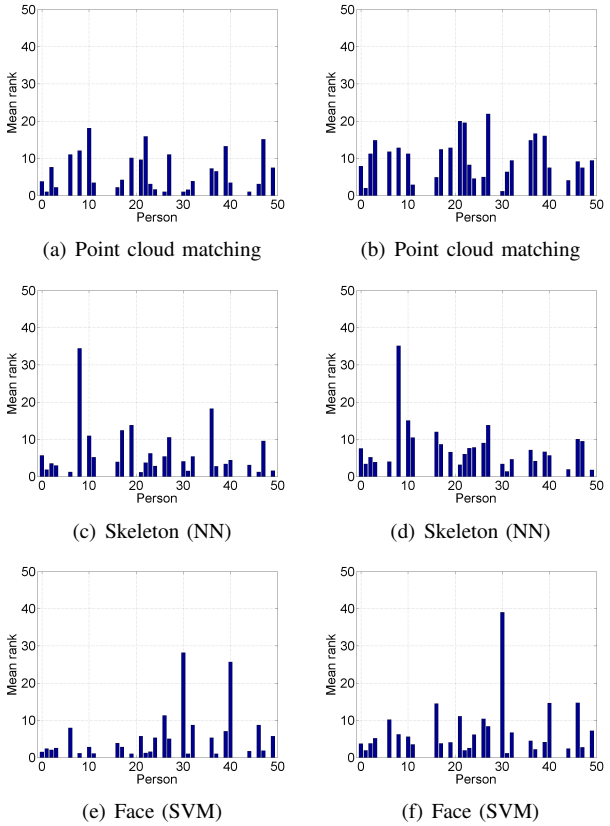
EVALUATION RESULTS OBTAINED IN CROSS VALIDATION AND WITH THE TESTING SETS OF THE *BIWI RGBD-ID Dataset*.

	Cross validation			Test - Still			Test - Walking		
	Rank-1	**nAUC**	**R1 Multi**	**Rank-1**	**nAUC**	**R1 Multi**	**Rank-1**	**nAUC**	**R1 Multi**
Point Cloud Matching	93.7%	99.6%	100%	32.5%	89.0%	42.9%	22.4%	81.6%	39.3%
Skeleton (NN)	80.5%	98.2%	100%	26.6%	89.7%	32.1%	21.1%	86.6%	39.3%
PCM+Skeleton	-	-	-	38.6%	91.8%	46.4%	27.4%	87.4%	42.9%
Face (SVM)	97.8%	99.4%	100%	44.0%	91.0%	57.1%	36.7%	87.6%	57.1%
Face+Skeleton (SVM)	98.4%	99.5%	100%	52.0%	93.7%	67.9%	43.9%	90.2%	67.9%

TABLE II

EVALUATION RESULTS OBTAINED IN CROSS VALIDATION AND WITH THE TESTING SETS OF THE *IAS-Lab RGBD-ID Dataset*.

	Cross validation			TestingA			TestingB		
	Rank-1	**nAUC**	**R1 Multi**	**Rank-1**	**nAUC**	**R1 Multi**	**Rank-1**	**nAUC**	**R1 Multi**
Point Cloud Matching	92.0%	98.7%	100%	28.6%	73.2%	63.6%	43.7%	81.7%	72.7%
Skeleton (NN)	86.2%	96.4%	100%	22.5%	73.8%	27.3%	55.5%	86.3%	81.8%
PCM+Skeleton	-	-	-	25.6%	75.5%	27.3%	63.3%	86.3%	81.8%

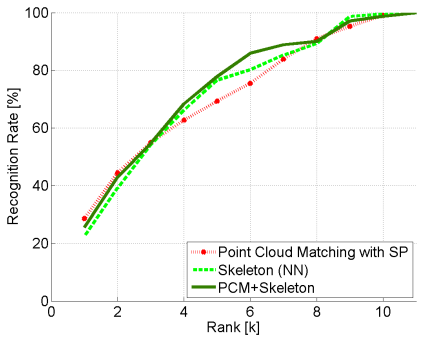


(a) Point cloud matching (b) Point cloud matching

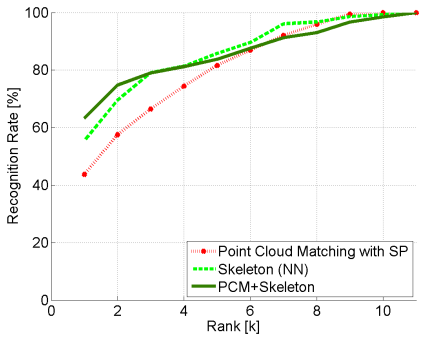
(c) Skeleton (NN) (d) Skeleton (NN)

(e) Face (SVM) (f) Face (SVM)

Fig. 6. Mean ranking histograms obtained with different techniques for every person of the *Still* (a-c) and *Walking* (d-f) test sets of the *BIWI RGBD-ID Dataset*.



(a) TestingA



(b) TestingB

Fig. 7. Cumulative Matching Characteristic Curves obtained with the main approaches described in this paper for the *IAS-Lab RGBD-ID Dataset*.

RGBD-ID Dataset, while all the numeric results are listed in Tab. II. It can be noticed that the absolute performance of the three methods considerably decreases for the testing set with people wearing different clothes (*TestingA*), but the point cloud matching technique provides a very good multi-frame rank-1 score (63.6%) which doubles that obtained with the skeleton descriptor (27.3%). For the *TestingB* set, the best results are again obtained by the combined approach, which obtained a rank-1 score of 63.3%, against 55.5% of the skeleton descriptor classification and 43.7% of the point cloud matching.

C. Runtime Performance

In Tab. III, the runtime of the single algorithms needed for the point cloud matching method of Sec. IV are reported. They refer to a C++ implementation running on a standard workstation with an Intel Core i5-3570k@3.40GHz processor. 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

TABLE III

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

takes 250ms for performing 50 comparisons. The overall frame rate is then of about 2.8fps, which suggests that this approach could be used in a real time scenario with further optimization and with a limited number of people in the database.

IX. CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

In this paper, we proposed an efficient method for composing 3D models of persons while moving freely. For overcoming the problem of the different poses a person can assume, we exploited the skeletal information provided by a skeletal tracking algorithm for warping persons point clouds to a standard pose, such that point clouds coming from different frames can be merged to compose a model. We showed how these models can be effectively used for the long-term re-identification task by means of a rigid comparison based on a ICP-like fitness score. We also compared the proposed technique with other state of the art approaches in terms of re-identification results on two newly created datasets designed for RGB-D re-identification. Moreover, we proposed a method for combining skeleton lengths and body shape information to further improve re-identification results. Experimental results show that shape information can be used to effectively re-identify subjects in a non-collaborative scenario, reaching performances near those of face recognition if PCM is combined with a classification of skeleton lengths. More accurate depth sensors and skeletal tracking algorithms would be helpful for obtaining more correct and realistic 3D models, so that a mobile robot could compose 3D models which could be used for re-identification by both robots and humans.

As future work, we will study the re-identification performance when varying the view points from which people are observed and in presence of partial failures of the transformation to standard pose. We also plan to develop algorithms for reducing the problem of holes and overlaps between parts in the point clouds transformed to standard pose.

REFERENCES

[1] A. K. Jain, S. C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?" *Proc. SPIE, Biometric Technology for Human Identification*, vol. 5404, pp. 561–572, 2004.

[2] D. Ober, S. Neugebauer, and P. Sallee, "Training and feature-reduction techniques for human identification using anthropometry," in *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, sept. 2010, pp. 1 –8.

[3] C. Velardo and J.-L. Dugelay, "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*, 2012, pp. 1–4.

[4] R. Satta, F. Pala, G. Fumera, and F. Roli, "Real-time appearance-based person re-identification over multiple Kinect cameras," in *VisApp*, 2013.

[5] B. I. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," in *First International Workshop on Re-Identification*, 2012.

[6] T. Leyvand, C. Meekhof, Y.-C. Wei, J. Sun, and B. Guo, "Kinect identity: Technology and experience," *Computer*, vol. 44, no. 4, pp. 94 –96, april 2011.

[7] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien, "Gait recognition with kinect," in *Proceedings of the First Workshop on Kinect in Pervasive Computing*, 2012.

[8] A. M. Bronstein, M. M. Bronstein, , and R. Kimmel, "Three-dimensional face recognition," *International Journal of Computer Vision*, vol. 64, pp. 5–30, 2005.

[9] A. M. Bronstein, M. M. Bronstein, and R. Kimmel, "Topology-invariant similarity of nonrigid shapes," *International Journal of Computer Vision*, vol. 81, pp. 281–301, March 2009.

[10] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, *et al.*, "Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera," in *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 2011, pp. 559–568.

[11] J. Chen, D. Bautembach, and S. Izadi, "Scalable real-time volumetric surface reconstruction," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 113, 2013.

[12] A. Weiss, D. Hirshberg, and M. Black, "Home 3d body scans from noisy image and range data," in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 1951–1958.

[13] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, September 2011, pp. 2044 –2049.

[14] W. Li, Z. Zhang, and Z. Liu, "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,*, June 2010.

[15] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2012.

[16] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from rgbd images," in *International Conference on Robotics and Automation*, 2012.

[17] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia, and B. Sankur, "The liris human activities dataset and the icpr 2012 human activities recognition and localization competition, Tech. Rep. RR-LIRIS-2012-004, March 2012.

[18] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "A comprehensive multimodal human action database," in *Proc. of the IEEE Workshop on Applications on Computer Vision*, 2013.

[19] P. J. Besl and N. McKay, "A method for registration of 3-d shapes," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 239–256, 1992.

[20] S. S. H. Jin and A. Yezzi, "Multi-view stereo reconstruction of dense shape and complex appearance," *Intl. J. of Computer Vision*, vol. 63, pp. 175–189, 2005.

[21] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European Conference on Computer Vision*, 2008, vol. 5302, pp. 262–275.

[22] P. A. Viola and M. J. Jones, "Robust real-time face detection," in *International Conference on Computer Vision*, 2001, p. 747.

[23] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, "Real-time facial feature detection using conditional regression forests," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[24] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, June 2008.