Eurographics Workshop on 3D Object Retrieval (2018)
A. Telea, T. Theoharis, and R. C. Veltkamp (Editors)

Person Re-Identification from Depth Cameras using Skeleton and 3D Face Data

P. Pala[1], L. Seidenari[1], S. Berretti[1] and A. Del Bimbo[1]

[1]University of Florence, Media Integration and Communication Center (MICC), Italy

Abstract

In the typical approach, person re-identification is performed using appearance in 2D still images or videos, thus invalidating any application in which a person may change dress across subsequent acquisitions. For example, this is a relevant scenario for home patient monitoring. Depth cameras enable person re-identification exploiting 3D information that captures biometric cues such as face and characteristic dimensions of the body. Unfortunately, face and skeleton quality is not always enough to grant a correct recognition from depth data. Both features are affected by the pose of the subject and the distance from the camera. In this paper, we propose a model to incorporate a robust skeleton representation with a highly discriminative face feature, weighting samples by their quality. Our method combining face and skeleton data improves rank-1 accuracy compared to individual cues especially on short realistic sequences.

CCS Concepts

•***Computing methodologies*** → ***Biometrics; Computer vision representations;*** *3D imaging;*

1. Introduction

The 3D scanning technologies substantially advanced in the last few years so that they can be used to capture geometric and visual data of an observed scene and its dynamics along time. The acquired depth and RGB frames are registered each other, thus boosting the potential of automatic analysis methods that can now easily detect and track people and their body parts as they move in the scene.

However, the technologies employed in current 3D dynamic scanning devices limit their field of view at a distance of few meters, with the quality of the sensed data degrading already at 2 meters distance. As a consequence, the tracking libraries released with such devices can track the target just if it is visible and sufficiently close to the sensor: if the moving target becomes too far from the sensor or it is no more in its field of view, the tracking is not possible. The ultimate result is that in the case a target observed in the past enters again the field of view of the camera, it is considered as a new one, loosing any relation between the two intervals of observation.

To exemplify a possible concrete scenario of application, let us consider the monitoring of a patient in a domestic environment as can be the case of elderly people or persons following a rehabilitation program at home. Suppose we want to monitor the long-term behaviour of the patient using one or multiple 3D sensors (like Kinect camera), each of them with a field of view constrained to a room or part of it. The ultimate goal of such a system could be the extraction of indices of position, movement, action, and behavior of the patient along days or weeks. This requires the correct identification of the monitored subject through subsequent temporal intervals, in which he/she is visible in the field of view of the cameras. Change in the appearance of the target subject as well as the presence of multiple persons should be also accounted for.

The task of person re-identification consists in recognizing an individual in different locations over a set of non-overlapping camera views [ZYH16]. Re-identification from depth images is facilitated by the joint face and body measurement. However, these measurements are far from accurate when using low cost sensors, such as Kinect. First, face imagery allows a face reconstruction via super-resolution only if a sufficient amount of views with enough resolution are available. On the other hand, skeleton is not always correctly estimated. Pose and distance may affect the accuracy of joints location estimation. Back and profile poses cause imprecise estimations. Moreover, when a subject is too close to the camera, many joints are occluded causing an almost total failure in the body feature computation. Figure 1 shows critical situations for both face and skeleton acquisitions.

Our model deals with these issues and allows us to perform re-identification accurately even if one of the two biometric cues is missing or inaccurately computed.

1.1. Our Contribution

In this paper, we present a model to gather and organize 3D data acquired by an RGB-D camera for the purpose of enabling long term re-identification of subjects observed by the camera. A cumu-

DOI: 10.2312/3dor.20181058

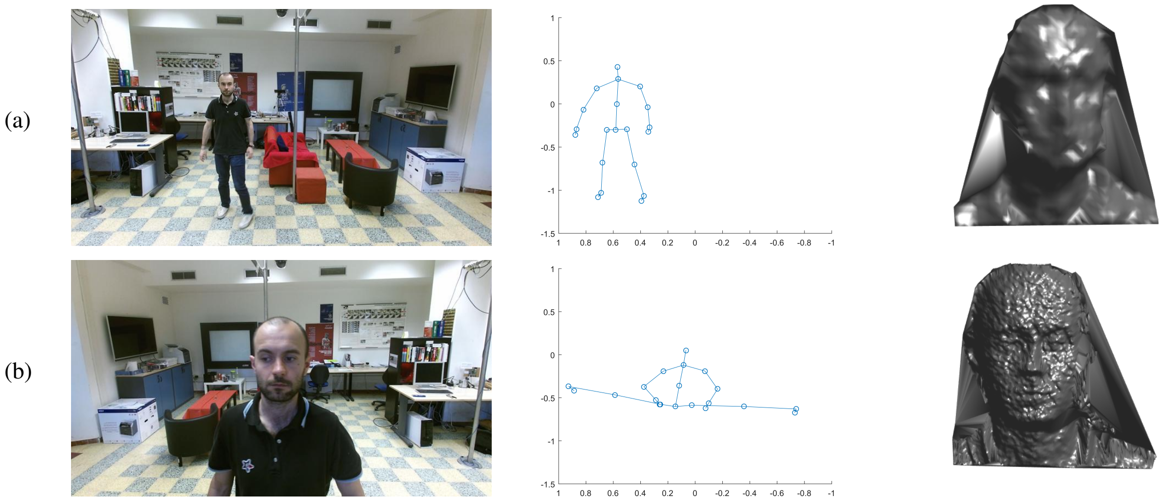


Figure 1: *Examples of skeleton and face mesh (Florence 3D Re-Id dataset): (a) for a far person (∼ 3m), the skeleton is estimated correctly, while the face mesh has very low quality; (b) for a close person (∼ 0.8m), leg joints are wrongly estimated, while the face mesh is noisy, but has high resolution.*

lated observed model is built for each subject, by retaining representative geometric and visual data of the subject from different viewpoints. The process of construction of the cumulated observed model is incremental allowing new observations of the subject to be incorporated in the model and replace old ones if the new observations are qualitatively better: in general, the subject distance to the camera and his/her speed of motion may affect the quality of acquired RGB and depth data. Data retained in the cumulated observed model are used to feed a 3D reconstruction module that outputs a 3D face of the subject to be used for re-identification.

To improve the robustness of the method, and its applicability, we also exploit skeletal features. Skeleton descriptors are also computed incrementally weighting their contribution according to a reliability measure. We propose a joint model, fusing both biometric cues that allows us to perform re-identification also in cases where one of the cues is not reliable.

To evaluate the proposed approach, we collected a dataset in our lab, which will be publicly released. Our dataset contains natural, unscripted, behavior of subjects acquired at various distances and poses.

The paper is organized as follows: In Section 2, previous work of person reidentification from depth data is summarized; Section 3 describes the model used to gather and organize multiple RGB and depth data coming from different observations of a subject; Section 4 expounds how these data are processed so as to compute a more accurate representation—compared to the accuracy of a single depth frame—of the geometry of the face of the subject. Such representation is used to enable subject re-identification; Section 5 describes how re-identification using the face geometry and the skeletal features is performed, and how these two can be fused together; finally, Section 6 reports the results of the evaluation of the proposed re-identification approach, also in comparison with alternative approaches; conclusions are given in Section 7.

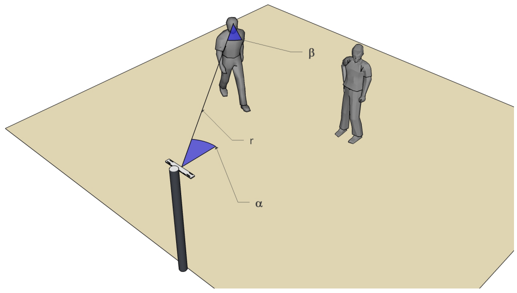


Figure 2: *The reference system. The subject position is accounted through the distance r measured along the ray connecting the camera to the subject, and the angles* α *and* β *formed by the ray and the viewing direction of, respectively, the camera and the subject.*

2. Related Work

Re-identification approaches have been developed first using 2D videos. Most of these 2D solutions rely on appearance-based only techniques, which assume that individuals do not change their clothing during the observation period [ZGX11, LMBD15]. This hypothesis constrains such re-identification methods to be applied under a limited temporal range.

Recently, the use of biometric features has been considered as a viable solution to overcome such limitations. In particular, there is an increasing interest in performing person re-identification using 3D data [BDTB18]. This idea has been first exploited using 3D soft biometric features. For example, Velardo and Dugelay [VD12] used anthropometric data obtained in a strongly supervised scenario, where a complete cooperation of the user is required to take manual measures of the body. However, in order to extend the applicability of re-identification systems to more practical scenarios, they should deal with subjects that do not explicitly cooperate with the system. This has been made possible thanks to the introduction of low cost 3D cameras capable of acquiring metric data of moving subjects in a dynamic way.

Some works approached the problem by combining appearance and depth data. Møgelmose et al. [MBM*13] presented a system where RGB, depth, and thermal data are combined for re-identification purposes. First, from each of the three modalities, some particular features are obtained: from RGB data, color information from different regions of the body is modeled; from depth data, different soft body biometrics are computed; and from thermal data, local structural information are extracted. Then, the three information types are combined in a joined classifier. Pala et al. [PSFR16] investigated whether the re-identification accuracy of clothing appearance descriptors can be improved by fusing them with anthropometric measures extracted from depth data, using RGB-D sensors, in unconstrained settings. Baltieri et al. [BVC14] proposed a re-identification framework, which exploits non-articulated 3D body models to spatially map appearance descriptors (color and gradient histograms) into the vertices of a regularly sampled 3D body surface. The matching and the shot integration steps are directly handled in the 3D body model, reducing the effects of occlusions, partial views or pose changes, which normally afflict 2D descriptors. A fast and effective model-to-image alignment is also proposed. It allows operation on common surveillance cameras or image collections. A comprehensive experimental evaluation is presented using the benchmark suite 3DPeS.

Several recent works exploited the opportunities given by depth sensors and performed person re-identification using soft-biometric cues. In [BCDB*12], Barbosa et al. presented a set of 3D soft-biometric cues that are gathered using RGB-D technology and being insensitive to appearance variations can be used for person re-identification. These include skeleton-based features (i.e., distances between joints of the skeleton, ratio between joint distances, and distances between joints and floor), and surface-based features (i.e., geodesic distances between joints computed on the reconstructed surface of the subject's 3D model). The joint use of these characteristics provides encouraging performances on a benchmark of 79 people that have been captured in different days and with different clothing. In [MBF*14], Munaro et al. proposed a method for creating 3D models of persons freely moving in front of a consumer depth sensor and show how they can be used for long-term person re-identification. To overcome the problem of the different poses a person can assume, the information provided by skeletal tracking algorithms is exploited 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. Karianakis et al. [KLCS17], targeted person re-identification from depth sensors such as Kinect. They explored the use of recurrent Deep Neural Networks for learning high-level shape information from low-resolution depth images. In order to tackle the small sample size problem, they introduced regularization and a hard temporal attention unit. The whole model can be trained end to end with a hybrid supervised loss. Wu et al. [WZL17] proposed to exploit depth information to provide more invariant body shape and skeleton information regardless of illumination and color change. They exploited depth voxel covariance descriptor and further propose a locally rotation invariant depth shape descriptor called Eigen-depth feature to describe pedestrian body shape. The effectiveness of the models was validated on publicly available depth pedestrian datasets.

However, none of the above methods considered the opportunity to combine together body and face depth data to improve re-identification.

3. Cumulated Observed Model of Body and Face

The setup of the system features a Kinect v2.0 camera mounted on a vertical pole at approximately 2 meters from the ground, and oriented so as to observe people entering and moving in a room (see the *reference system* in Fig. 2). Using the Kinect SDK, the camera outputs RGB and depth frames as well as the 3D coordinates and orientation of the skeleton joints, for up to 6 persons. These data are processed to compute the position and orientation of a generic subject within the field of view of the camera in terms of the *radial distance* r, the *azimuthal angle* α, and the *yaw angle* β (see Fig. 2). Pitch and roll angles, although provided by the SDK are presently not considered.

Values of (r,α,β) are discretized so as to represent the position and orientation of a generic subject with respect to the camera by using the the triple (i,j,k) to index one among N_c possible configurations. Given the observation (r,α,β) representing the position and orientation of a generic subject with respect to the camera, quantized observed configuration indexes (i_o,j_o,k_o) are computed as:

$$\begin{cases} i_o = \arg\min_i |r_i - r| \,, & i = \{1,\ldots,N_r\} \\ j_o = \arg\min_j |\alpha_j - \alpha| \,, & j = \{1,\ldots,N_\alpha\} \\ k_o = \arg\min_k |\beta_k - \beta| \,, & k = \{1,\ldots,N_\beta\} \,. \end{cases} \tag{1}$$

For a generic observation, a confidence measure is estimated to express the presence of out of focus artifacts in the RGB data caused by subject motion or inadequate lighting. In this way, a new observation with quantized configuration indexes (i_o,j_o,k_o) replaces the previous observation with the same quantized configuration indexes only if the confidence of the new observation is greater than the confidence of the previous one. Figure 3 shows an example of the observations retained after tracking a subject who wandered in front of the camera for some time.

In addition to this multiview representation of the face, our Cumulative Observation Model (COM) retains a representation of the skeleton of the observed person. This is achieved by computing an exponential moving average measure of the distance between some pairs of body joints.

By adopting an exponential weighted moving average measure of the body parts, the accuracy of the skeleton based representation of the observed person increases with the duration of the observation. This enables the use of these data to complement facial data and increase the accuracy of re-identification.

We weigh each skeletal descriptor according to our reliability function:

$$rel(s) = \frac{|\mathcal{J}_T|}{|\mathcal{J}|} + \frac{1}{2}\cdot(1-\mathbf{z}\cdot\mathbf{v}) + \frac{||\text{head}-\text{head}_{gp}||}{H_{geo}} \,. \tag{2}$$

The reliability function $rel(s)$ has three terms:

- $\frac{|\mathcal{J}_T|}{|\mathcal{J}|}$ takes into account the reliability of the joint tracking by computing the ratio of tracked joints $j \in \mathcal{J}_T$ with respect to the whole joint set $\mathcal{J}$;
- $\frac{1}{2} \cdot (1 - \mathbf{z} \cdot \mathbf{v})$ evaluates the body pose, where $\mathbf{z}$ is the vector indicating the z axis in the camera reference and $\mathbf{v}$ is the vector perpendicular to the plane estimated from torso joints;
- $\frac{||\text{head}-\text{head}_{gp}||}{H_{geo}}$ evaluates how *erected* a subject pose is. H_{geo} is the geodesic height, defined as:

$$\begin{aligned} H_{geo} = \; & ||\text{head} - \text{neck}|| + ||\text{spine-mid} - \text{spine-base}|| + \\ & \tfrac{1}{2}(||\text{left-hip} - \text{left-knee}|| + ||\text{lknee} - \text{lankle}|| + \\ & ||\text{rhip} - \text{rknee}|| + ||\text{rknee} - \text{rankle}||) \; , \end{aligned}$$

where head$_{gp}$ is the projection of the head onto the ground plane. Note that in computing H_{geo}, we average on the leg lengths for improved accuracy. Considering a skeleton descriptor at frame t, s_t, we compute the cumulated observation for a sequence of skeletons $\mathcal{S}$ as:

$$s^* = \sum_{s_t \in \mathcal{S}} d_\alpha(t) \cdot rel(s_t) \cdot s_t \; , \qquad (3)$$

where $d_\alpha(t) = \exp\left(\frac{t}{\tau}\right)$ is an exponential decay term that weights decreasingly the relevance of descriptors s_t (details on the actual form of the descriptor are given in Section 5).

4. Reconstructed Face Model

Observations retained from different viewpoints by the COM are used to build a 3D model of the face of the subject using a 3D super-resolution approach, developing on the model proposed in [BPD14].

Each range image retained by the COM is converted into a point cloud, and information about the acquisition radius, azimuth and yaw angles are used to roughly align the different point clouds to a common (X,Y,Z) reference system. The Iterative Closest Point (ICP) algorithm [RL01] is then used for fine registration of the point clouds with respect to each other. Once all the point clouds are registered and aligned to a common reference system, estimation of the face surface is operated by fitting a mean face model to the data (points of the clouds). This is performed in two steps: mean face model alignment, and mean face model warping. The ICP algorithm is used for alignment, whereas warping is accomplished by updating the coordinates of each vertex of the mean face model based on the spatial distribution of the closest points of the cloud. The deformable face model proposed in [FLBD17] is used as mean face model.

Formally, considering one generic vertex $\vec{v} = (v_x, v_y, v_z)$ of the mean face model, the subset of the point cloud (PC) composed of points within a range Δ from the vertex is considered:

$$\mathcal{S}(\vec{v}) = \{\vec{x} \in PC| \, ||\vec{v} - \vec{x}|| < \Delta\} \; . \qquad (4)$$

Each point $\vec{x}_i \in \mathcal{S}(\vec{v})$ is assigned a weight w_i accounting for its distance to $\vec{v}$. Eventually, the coordinates of $\vec{v}$ are updated through the following expression:

$$\vec{v} = \frac{\sum w_i \vec{x}_i}{\sum w_i} \; . \qquad (5)$$

Figure 4 shows two sample facial point clouds retained by the COM, the cumulated facial point cloud obtained by registering all the retained point clouds, the mean face model before and after the warping process.

5. Re-identification based on Face and Body Part Geometry

Re-identification based on face geometry operates by reconstructing a 3D face model of each observed person and matching this probe against a gallery set composed of reconstructed 3D face models of previously observed persons. In the case a match is found the person is reidentified. Description and matching of gallery and probe models is obtained according to the approach proposed in [BDP13] that is based on the extraction and comparison of local features of the face. First, SIFT keypoints of the face are detected from the depth image of the face, and a subset of them is retained by applying a hierarchical clustering. In this way, a cluster of keypoints with similar position and SIFT descriptors is substituted by a "representative keypoint", thus reducing the overall number of keypoints. Then, the relational information between representative keypoints is captured by measuring how the face geometry changes along the surface path connecting pairs of keypoints. By sectioning the face through a plane passing from the two keypoints and orthogonal to the surface a *facial curve* is extracted. Face similarity is evaluated by finding correspondences between keypoints of probe and gallery scans, and matching the facial curves across the inlier pairs of matching keypoints. The approach revealed good performance across different datasets and also in the case of partial face matching. This provides the 3D face recognition approach with the required robustness to manage our scenario.

For re-identification based on body part geometry, due to the fact arms and legs are often wrongly located by Kinect, we only rely on features computed from the torso. Indeed, knees and hands have the lowest recognition rate [SGF*13]. We use neck, spine, shoulders and hips, and specifically we compute the following features using Euclidean distances:

$$\begin{aligned} s^{ns} &= ||\text{neck} - \text{spine-mid}|| \\ s^{mb} &= ||\text{spine-mid} - \text{spine-base}|| \\ s^{nls} &= ||\text{neck} - \text{lshould}|| \\ s^{nrs} &= ||\text{neck} - \text{lshould}|| \\ s^{lhb} &= ||\text{lhip} - \text{spine-base}|| \\ s^{rhb} &= ||\text{rhip} - \text{spine-base}|| \\ s^{mls} &= ||\text{spine-mid} - \text{lshould}|| \\ s^{mrs} &= ||\text{spine-mid} - \text{rshould}|| \; . \end{aligned}$$

For a skeleton at time t, $\mathcal{S}_t$, we define the 9-dimensional descriptor:

$$s_t = [s_t^{ns} s_t^{mb} s_t^{nls} s_t^{nrs} s_t^{lhb} s_t^{rhb} s_t^{mls} s_t^{mrs}] \; . \qquad (6)$$

Re-identification based on skeletal features is thus performed by sorting distances of probe cumulated skeleton descriptor with previously acquired cumulated descriptors of candidates.

Finally, we can combine face and body re-identification as follows. Let us consider a sequence as a set $\mathcal{T}$ of ordered tuples

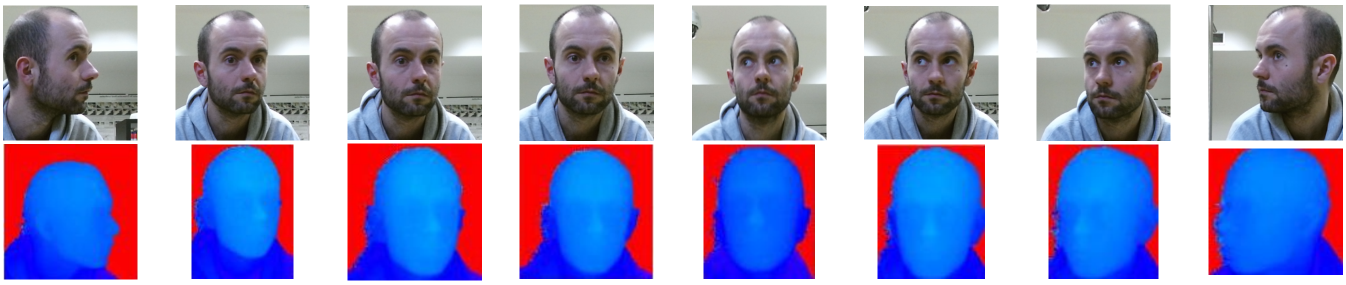


Figure 3: *Example of representative views of a subject retained by the cumulative observation model (Florence 3D Re-Id dataset).*

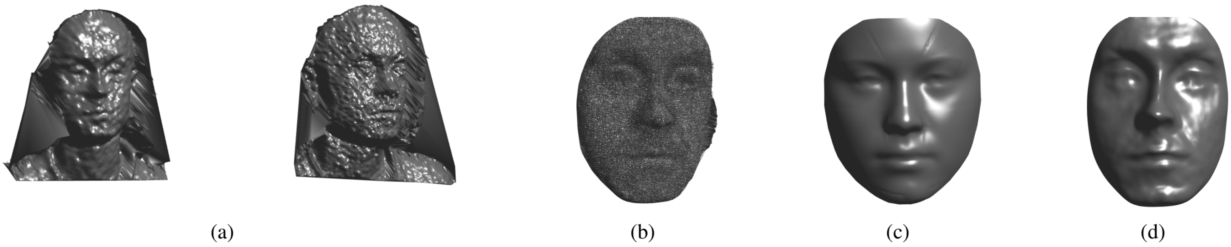


Figure 4: *Construction of the face model using observations from multiple viewpoints. (a) Two sample facial point clouds retained by the COM; (b) the cumulated facial point cloud obtained by registering all the retained point clouds; (c) the mean face model before, and (d) after the warping process.*

$\mathbf{t}_t : \langle \mathbf{f}_t, \mathbf{s}_t \rangle$, where $\mathbf{f}_t$ is a face crop from the depth image and $\mathbf{s}_t$ is a set of skeletal joint feature as defined above. Applying the COM to $\mathcal{T}$, we can obtain the cumulated model for face $\mathbf{f}$ and skeleton $\mathbf{s}$. To perform re-identification, let us consider a probe $\mathbf{t}_p := \langle \mathbf{f}_p, \mathbf{s}_p \rangle$. Re-identification is the task of sorting identities $\mathcal{I}$ in the gallery $\mathcal{G}$ by similarity with probe $\mathbf{t}_p$. We compute a distance for each identity $\mathcal{I}$ accumulating distances of every subsequence in the gallery:

$$D^f(\mathcal{I}, \mathbf{f}_p) = \sum_{i \in \mathcal{I}} d(\mathbf{f}_i, \mathbf{f}_p) \cdot \text{rank}^f(i) \; , \tag{7}$$

and for skeletons

$$D^s(\mathcal{I}, \mathbf{s}_p) = \sum_{i \in \mathcal{I}} d(\mathbf{s}_i, \mathbf{s}_p) \cdot \text{rank}^s(i) \; , \tag{8}$$

where i is a sample of identity $\mathcal{I}$, $\text{rank}^f(i)$ and $\text{rank}^s(i)$ are rank of sample i according to face and skeleton feature distance.

We compute the final identity ranking using:

$$D(\mathcal{I}, \mathbf{t}_p) = \alpha D^f(\mathcal{I}) + (1-\alpha) D^s(\mathcal{I}) \; , \tag{9}$$

where we set $\alpha = 0.6$ considering the better performance of face alone (this value has been determined on a preliminary set of experiments on a small set of training data).

6. Experimental Results

Re-identification experiments have been performed separately for face and skeleton, and for their fusion. In the following, we first summarize the dataset used, then report on the obtained results.

6.1. Dataset

We collected "Florence 3D Re-Id", a novel dataset of people performing natural gestures at varying distances from the sensor. Many previously collected datasets picture unnatural motions, such as standing still in front of the camera, or walking in circle. We instruct subjects to move in front of the sensor varying their distance, in order to capture biometric cues in different conditions. We also allow and encourage subjects to perform any task they are willing to do, such as reading their watch, interacting with a smart-phone or answering a call. All these actions are performed without any time line or choreography. Figure 1 shows two sample frames from our dataset, highlighting challenging situations that can happen in the case either the quality of the acquisition for skeleton or face data are low. So, our dataset includes strong variations in terms of distance from the sensor, pose, and occlusions.

We record three separate sequences for each of the 16 subjects included in the dataset. The first two sequences contain different behaviors performed standing. The third sequence of each subject pictures a sit-down and stand-up sequence in order to analyze the criticality of skeletal representation for non-standing poses. In particular, in this latter case, the joints estimation provided by the Kinect camera is more critical due to self-occlusions. Potentially, more stable solutions for occluded joints estimation could be used [RGL15]. We collect depth frames at a 512 × 424 resolution (Kinect 2 standard), and the skeleton estimation with joint state (tracked/estimated). We also collect, but do not use in this work, face landmarks and the 3D face model fitted by the Microsoft SDK.

The dataset is comprised of 39,315 frames. Skeletons are acquired in 17,982 frames, while faces are captured at a distance suitable for reconstruction ($0.5 \sim 1.5m$) in 2,471 frames.

6.2. Face Re-identification Results

In this experiment, we performed re-identification by using the models of the face reconstructed using full sequences and subsequences with 300, 200, and 100 frames, respectively. In this way, we can evaluate the behavior of our model on sequences with different number of frames, and observe how this impacts on the selection of "good" frames for reconstruction. This behavior can be visually appreciated in Fig. 5, where some reconstruction examples using the full sequence, and sequences with 300, 200 and 100 frames are reported. It can be noted, there is quite a large variability in the quality of the reconstructed models in the case only part of the sequence is used, and in general the perceived visual quality improves with the number of frames.

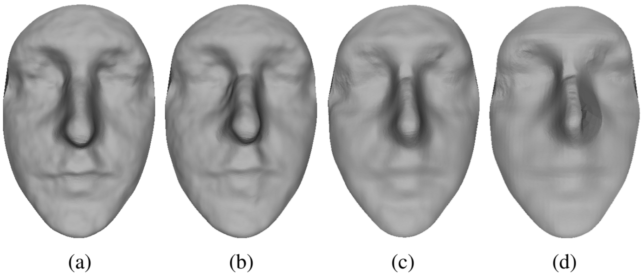


Figure 5: *Models reconstructed for one subject using: (a) full sequence; (b) 300 frames; (c) 200 frames; (d) 100 frames.*

For comparing reconstructed face models, the face description and matching approach proposed in [BDP13] has been used. Results are reported in Table 1. Quite evidently it emerges the performance drop in using full and partial sequences.

Table 1: *Re-identification rate (RR) using face models reconstructed on sequences with different number of frames.*

Sequence length	#probes	RR
Full sequences	32	93.8%
sub-sequences 300	75	65.3%
sub-sequences 200	87	56.3%
sub-sequences 100	106	56.6%

6.3. Body Re-identification Results

We run a set of experiments to evaluate our cumulated model and our set of features for re-identification. We vary the timeframe over which recognition is performed. We show in Table 2 the difference between the weighted and unweighted model. The use of Eq. (2) to weight skeleton features allows better recognition rate. Clearly, the larger the set of skeletons influencing the final descriptor the better the recognition. On full sequences weighting skeleton quality allows an improvement of 7% in recognition accuracy, which is much more than for shorter sequences. This is motivated by the fact that in longer sequences there is a higher chance of finding highly unreliable skeletons, which if unweighted will drastically worsen the performance.

Table 2: *Rank-1 recognition rate varying timeframe constant τ, using Eq. (2) (weighted) or not (unweighted).*

Sequence length	weighted	unweighted
Full sequences	**41.7**	34.7
sub-sequences 300	**31.3**	30.2
sub-sequences 200	**31.0**	30.1
sub-sequences 100	**28.7**	27.9

6.4. Evaluation of the Fusion between Face and Body

Finally, we report the CMC curves on sub-sequences of different length evaluating our fused model exploiting skeleton and face re-identification jointly. In Fig. 6, we report CMC for different subsequence length. In the ideal case of full sequences, the use of skeleton does not add much to the almost perfect recognition we obtain from super-resolved faces, with a rank-1 recognition rate of 93.8%. In more realistic scenarios, when less frames are available, it can be seen that the fusion of the two features is extremely valuable. Indeed, faces have always a better rank-1 recognition rate, but the fusion model scores always higher than face and skeleton alone, raising rank-1 accuracy too.

7. Conclusions

In this paper, we presented a method for person re-identification from 3D sensors. We showed how reconstructed faces with increased resolution can be derived from low-resolution depth frames, and as the resulting cumulated observed model can be used to recognize people very effectively. We also presented an analogous strategy to cumulate observations of body skeletons. Recognition accuracy using skeletal data is less effective when compared to that obtainable from face data, but it is more applicable at a distance. Finally, we evidenced that combining face and skeleton outperforms both single cue methods on short realistic sequences.

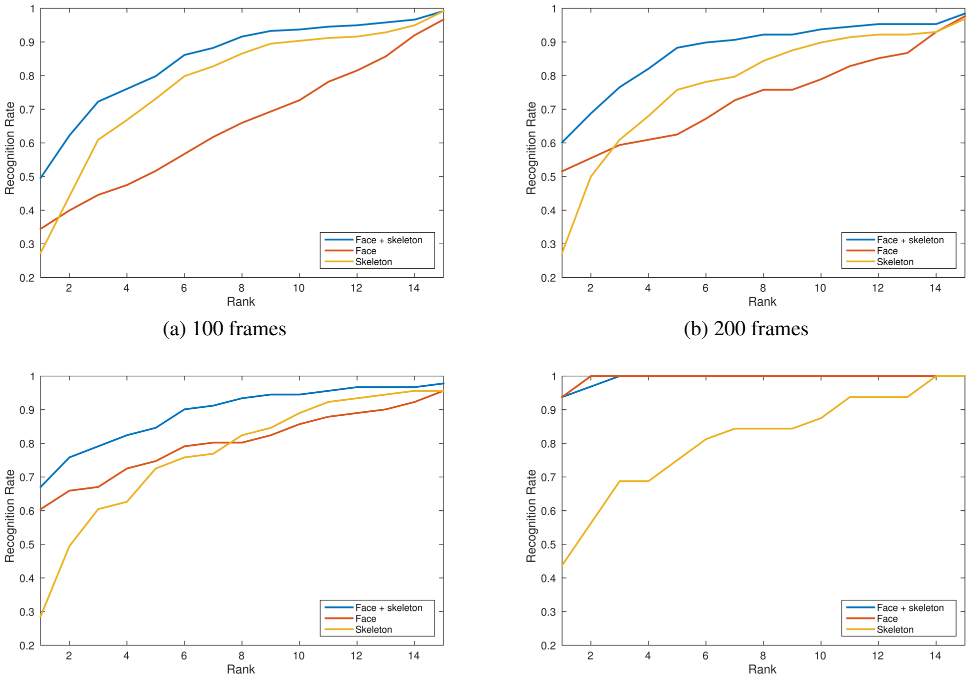
8. Acknowledgment

This research is funded by the "Ente Cassa di Risparmio Firenze", under the project CAMPO (Caratterizzazione Multivista di Persone Osservate da Sensori RGB-D per Ri-identificazione).

References

[BCDB*12] BARBOSA B. I., CRISTANI M., DEL BUE A., BAZZANI L., MURINO V.: Re-identification with RGB-D sensors. In *Int. Workshop on Re-Identification, in European Conference on Computer Vision (ECCV) Workshops and Demonstrators* (Florence, Italy, Oct. 2012), Springer, (Ed.), vol. LNCS 7583, pp. 433–442. 3

[BDP13] BERRETTI S., DEL BIMBO A., PALA P.: Sparse matching of salient facial curves for recognition of 3D faces with missing parts. *IEEE Trans. on Information Forensics and Security 8*, 2 (Feb. 2013), 374–389. 4, 6



(a) 100 frames

(b) 200 frames

(c) 300 frames

(d) all frames

Figure 6: *CMC for fusion model on sequences with 100, 200, 300 frames and full sequences (all frames). The fusion model helps especially on short sub-sequences.*

[BDTB18] Berretti S., Daoudi M., Turaga P., Basu A.: Representation, analysis and recognition of 3d humans: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications 14*, 1s (March 2018), 1–35. 2

[BPD14] Berretti S., Pala P., Del Bimbo A.: Face recognition by super-resolved 3D models from consumer depth cameras. *IEEE Trans. on Information Forensics And Security 9*, 9 (Sept. 2014), 1436–1449. 4

[BVC14] Baltieri D., Vezzani R., Cucchiara R.: Mapping appearance descriptors on 3D body models for people re-identification. *International Journal of Computer Vision 111*, 3 (2014), 345–364. 3

[FLBD17] Ferrari C., Lisanti G., Berretti S., Del Bimbo A.: A dictionary learning-based 3d morphable shape model. *IEEE Trans. on Multimedia 19*, 12 (Dec 2017), 2666–2679. doi:10.1109/TMM.2017.2707341. 4

[KLCS17] Karianakis N., Liu Z., Chen Y., Soatto S.: Person depth reid: Robust person re-identification with commodity depth sensors. *CoRR abs/1705.09882* (2017). URL: http://arxiv.org/abs/1705.09882, arXiv:1705.09882. 3

[LMBD15] Lisanti G., Masi I., Bagdanov A., Del Bimbo A.: Person re-identification by iterative re-weighted sparse ranking. *IEEE Trans. on Pattern Analysis and Machine Intelligence 37*, 8 (Aug 2015), 1629–1642. 2

[MBF*14] Munaro M., Basso A., Fossati A., Gool L. V., Menegatti E.: 3D Reconstruction of Freely Moving Persons for Re-Identification with a Depth Sensor. In *IEEE Int. Conf. on Robotics and Automation (ICRA)* (Hong-Kong, May 2014), pp. 4512–4519. 3

[MBM*13] Møgelmose A., Bahnsen C., Moeslund T. B., Clapes A., Escalera S.: Tri-modal person re-identification with rgb, depth and thermal features. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2013), pp. 301–307. doi:10.1109/CVPRW.2013.52. 3

[PSFR16] Pala F., Satta R., Fumera G., Roli F.: Multimodal person re-identification using RGB-D cameras. *IEEE Trans. on Circuits and Systems for Video Technology 26*, 4 (April 2016), 788–799. 3

[RGL15] Rafi U., Gall J., Leibe B.: A semantic occlusion model for human pose estimation from a single depth image. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)* (June 2015), pp. 67–74. 5

[RL01] Rusinkiewicz S., Levoy M.: Efficient variants of the ICP algorithm. In *Proc. Int. Conf. on 3D Digital Imaging and Modeling (3DIM)* (Quebec City, Canada, May 2001), pp. 145–152. 4

[SGF*13] Shotton J., Girshick R., Fitzgibbon A., Sharp T., Cook M., Finocchio M., Moore R., Kohli P., Criminisi A., Kipman A., et al.: Efficient human pose estimation from single depth images. *IEEE Trans. on Pattern Analysis and Machine Intelligence 35*, 12 (2013), 2821–2840. 4

[VD12] Velardo C., Dugelay J.: Improving identification by pruning: A case study on face recognition and body soft biometric. In *Int. Work. on Image Analysis for Multimedia Interactive Services (WIAMIS)* (Dublin, Ireland, May 2012), pp. 1–4. 2

[WZL17] Wu A., Zheng W. S., Lai J. H.: Robust depth-based person re-identification. *IEEE Trans. on Image Processing 26*, 6 (June 2017), 2588–2603. doi:10.1109/TIP.2017.2675201. 3

[ZGX11] Zheng W.-S., Gong S., Xiang T.: Person re-identification by probabilistic relative distance comparison. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (Colorado Springs, CO, USA, June 2011), pp. 649–656. 2

[ZYH16] Zheng L., Yang Y., Hauptmann A. G.: Person re-identification: Past, present and future. *CoRR abs/1610.02984* (2016). URL: http://arxiv.org/abs/1610.02984, arXiv:1610.02984. 1