Eurographics Workshop on 3D Object Retrieval (2016)
A. Ferreira, A. Giachetti, and D. Giorgi (Editors)

Shrec'16 Track: Retrieval of Human Subjects from Depth Sensor Data

A.Giachetti[1], F.Fornasa[1], F.Parezzan[1], A.Saletti[1], L.Zambaldo[1], L.Zanini[1],

F. Achilles[2,3], A. E. Ichim[4], F. Tombari[2,4], N. Navab[2,6], S.Velasco-Forero[7]

[1]Department of Computer Science University of Verona, Italy - Track organizers
[2] Computer Aided Medical Procedures, Technische Universitat Munchen, Germany
[3] Department of Neurology, Ludwig-Maximilians-University of Munich, Germany
[4] Graphics and Geometry Laboratory, École Polytechnique Fédérale de Lausanne, Switzerland
[5] DISI, University of Bologna, Italy
[6] Computer Aided Medical Procedures, Johns-Hopkins-University, U.S.A.
[7]CMM / Mines-Paris Tech

Abstract

In this paper we report the results of the SHREC 2016 contest on "Retrieval of human subjects from depth sensor data". The proposed task was created in order to verify the possibility of retrieving models of query human subjects from single shots of depth sensors, using shape information only. Depth acquisition of different subjects were realized under different illumination conditions, using different clothes and in three different poses. The resulting point clouds of the partial body shape acquisitions were segmented and coupled with the skeleton provided by the OpenNI software and provided to the participants together with derived triangulated meshes. No color information was provided. Retrieval scores of the different methods proposed were estimated on the submitted dissimilarity matrices and the influence of the different acquisition conditions on the algorithms were also analyzed. Results obtained by the participants and by the baseline methods demonstrated that the proposed task is, as expected, quite difficult, especially due the partiality of the shape information and the poor accuracy of the estimated skeleton, but give useful insights on potential strategies that can be applied in similar retrieval procedures and derived practical applications.

Categories and Subject Descriptors (according to ACM CCS): I.4.8 [IMAGE PROCESSING AND COMPUTER VISION]: Scene Analysis—Shape

1. Introduction

One of the most interesting application of non-rigid shape retrieval is certainly the one related to human subjects re-identification, that can have relevant applications, for example in security and surveillance. Recent papers [PSR*14,WMKS*15] demonstrated that quite good retrieval performances can be obtained for human bodies acquired with a whole body scanner. However, many real world applications could not be based on the acquisition of complete high resolution models of the subjects, but rather partial scans of low quality, like, for example, those that can be acquired with a depth sensor. Low cost depth sensors like Microsoft Kinect, Intel Realsense, etc., are now widely available and, despite some technical limitations due to the technologies used (IR structured light or Time of Flight sensing), they can now be used similarly to conventional cameras in surveillance and monitoring applications.

This fact suggested us to test geometry-based shape retrieval methods on the practical and extremely challenging task of retrieving from a database partial models of a human subject acquired with a low end depth sensor given an example. Human reidentification depth sensor datasets have already been proposed in the Computer Vision community [BCDB*12,MFB*14] and it is surely interesting to evaluate the contribution of purely geometric methods on this kind of applicative tasks.

2. Data acquisition and proposed task

Models have been created by placing a depth sensor (Asus Xtion Live Pro) in a position simulating typical a surveillance acquisition setup, with the depth camera placed in an elevated position (about 2.2m. from the floor), looking down with an angle of 22 degrees with respect to the horizontal plane (see Figure 1).

DOI: 10.2312/3dor.20161086

With this setup, we acquired depth maps of a group of subjects in three different poses with three different clothing (two scans with different coats and one without) and two different illumination conditions (natural light and artificial light), for a total of 18 scans for each subject. For each subject we recorded the acquired point clouds and the corresponding skeletons provided by the OpenNI functions (Figure 2). Rough point clouds have been processed in order to transform the coordinate system in order to have the normal to the floor plane along the y direction and subject x-z position approximately in the origin. Finally clouds' points belonging to the floor and to the environment were removed and a smoothing procedure based on the original structured point cloud connectivity was applied. Participants were finally provided with the rough and smoothed point clouds as ASCII .ply files, the skeleton file with coordinates of 15 nodes (HEAD, NECK, LEFT SHOULDER, LEFT ELBOW, LEFT HAND, RIGHT SHOULDER, RIGHT ELBOW, RIGHT HAND, TORSO, LEFT HIP, LEFT KNEE, LEFT FOOT, RIGHT HIP, RIGHT KNEE, RIGHT FOOT) and connectivity in .off format, and a triangulated mesh obtained with Meshlab [CCC*08] implementation of the ball pivoting algorithm [BMR*99] (Figure 4).

We acquired a total of 50 subjects (20-25 years old males and females), subdividing then the dataset in a training set of 180 models of 10 subjects and a test dataset with 720 scans of 40 different subjects. The training data set was also provided with label information and could in principle used to set algorithm parameters or train supervised methods.

Figure 3 shows examples of (cleaned) point clouds showing differences in the models of the same subject in the different acquisition conditions (lights, pose, clothing).

Participants were finally asked to send up to three dissimilarity matrices evaluated distances between all the shapes in the test set.

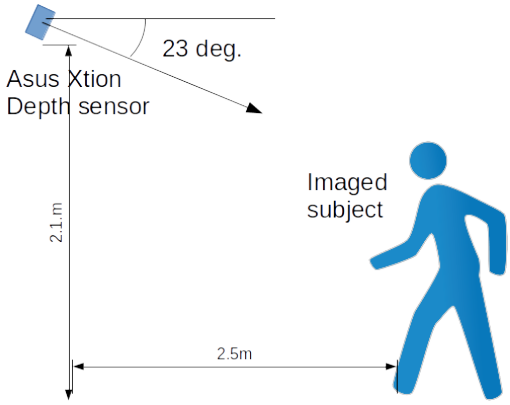

Figure 1: *Acquisition setup*

3. Evaluation

The retrieval performance of baseline and participants' methods were evaluated according to the classical measures used in [SMKF04], e.g. Nearest Neighbor (NN), First Tier (FT), Second Tier (ST), e-measure (E) and Discounted Cumualted Gain (DCG).

Figure 2: *Example of RGB image (left) depth map (center) and skeleton (right) obtained by the sensor.*

Furthermore, Precision-Recall plots have been analyzed and from the PR curves the Mean Average Precision (MAP), e.g. the average of all precision values computed for each subject in the retrieved list was estimated. An analysis of the effects of pose and clothing on the retrieval scores of the different methods has then been performed, and will be discussed in Section 6.

4. Baseline methods by A.Giachetti, F.Fornasa, F.Parezzan, L.Zanini

As basic method to characterize shape we propose the following descriptors:

Lengths of Skeletal Segments (LSS): we only used the lengths of the 14 skeletal segments as shape descriptors and used Euclidean distances to evaluate the distance matrix.

Statistics on Shape Points Clusters (SPC): we clustered the smoothed cloud points according to the closest skeletal segment and computed statistics on the point distribution. We used as descriptor components the average distance of the points from the segment (SPC Mean), the standard deviation of the distance (SPC STD) and the normalized concatenation of Mean and Standard deviation (SPC M+S). Euclidean distance was used to compute dissimilarity. As in some poses point clusters related to some skeletal segments are empty, we replaced the differences of the related components with the average of the well defined differences.

5. Participants and methods proposed

Only two groups participated to the contest. Unfortunately the difficulty of the task makes probably too difficult the use of standard geometry processing descriptors used in watertight mesh retrieval and the use of shape only information is not usual in the Computer Vision community. We received results and methods from the "TEAM TUM-EPFL", composed by F. Achilles, A-E. Ichim, F. Tombari, and N. Navab. and from Santiago Velasco-Forero (CMM MINES Paristech). In the following we describe the methods proposed by the participants.

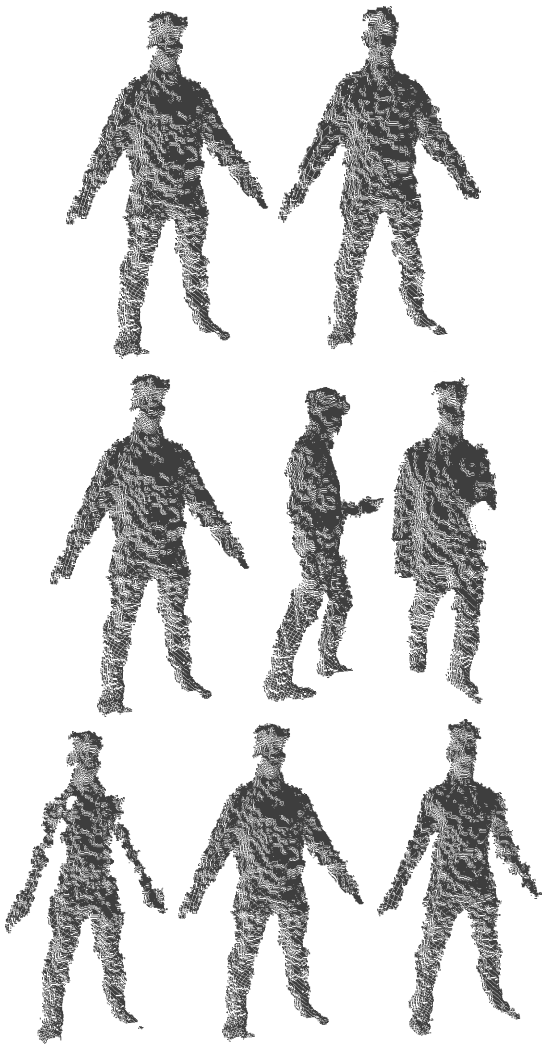

Figure 3: *Top row: same subject acquired with different illumination conditions. Middle row: same subject in the three poses. Bottom row: same subject with three different coats.*

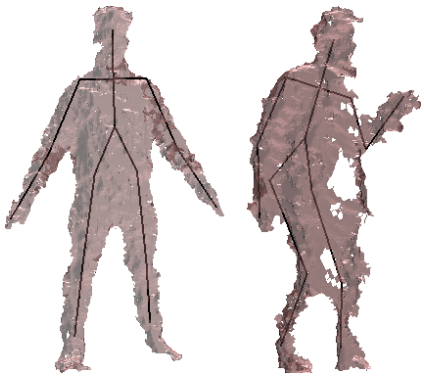

Figure 4: *Examples of triangulated meshes provided to participants together with the registered skeletal segments.*

Figure 5: `Mesh` *data an its correspondent* `Ske` *(In blue). Results of the operator* $Proj_{\texttt{Mesh}}(\texttt{Ske})$ *are shown in red.*

5.1. Step in Depth+Mesh+Skeleton classification by S. Velasco-Forero

For a given couple of (mesh,skeleton) denoted by $(\texttt{Mesh}, \texttt{Ske})$, tlhe goal was to produce a descriptor capturing the interaction between skeleton points in `Ske` and the mesh data `Mesh` calculated from depth information. Due to low quality of the mesh, in many cases `Mesh` contains holes and isolated part. However, `Ske` is complete by included hidden parts by symmetry. Thus, in this method, the author has projected `Ske` on `Mesh` and used the upper part of all-pairs distance as descriptor.

Here is a detailed description of the proposed algorithm.

1. For each $v \in \texttt{Ske}$ compute its closest point in the mesh, i.e.
$$\text{Proj}_{\texttt{Mesh}}(v) = \arg\min_{w \in \texttt{Mesh}} ||v-w||_2^2 \quad (1)$$
2. For every pair of points in the skeleton compute the pairwise distance, i.e. $\mathbf{D}(i,j) = ||\text{Proj}_{\texttt{Mesh}}(v_i) - \text{Proj}_{\texttt{Mesh}}(v_j)||_2^2$, for all $i,j = 1,\ldots,|\texttt{Ske}|$.
3. As $\mathbf{D}$ is symmetric, authors used $\textbf{vec}(\texttt{Upper}(\mathbf{D}))$ as descriptor, where `Upper` denotes the upper-triangular matrix capturing all the values above the diagonal and **vec** is the vectorization operator.

Three measures of similarities have been then considered via L_p-norm distances, as follows,

$$\text{Dist}(a,b) = \sum_{i=1,j>i}^{|\texttt{Ske}|} (|\texttt{Upper}(\mathbf{D_a})(i,j) - \texttt{Upper}(\mathbf{D_b})(i,j)|^p)^{1/p} \quad (2)$$

with $p = 1, 2, 1.5$. The first two cases are well-known as city-block(Manhattan) and Euclidean distances and are referred in results as $DMS-C$ and $DMS-E$. The third is referred to as $DMS-M$ (method using Minkowski distance).

5.2. Shape parameter estimation using a ConvNet and confidence weighting by F. Achilles, A. E. Ichim, F. Tombari, N. Navab

In order to retrieve human shapes from depth data, the group parameterized human mesh models using blend shapes [PSS99, BRM08]. To build up the database, the mesh model is rendered in several depth images with varying viewpoint- and shape parameters. Additionally, motion capture sequences were used to animate the model and hence induce robustness with respect to pose changes.

The connection between depth input and shape parameters were learned using a convolutional neural network, which jointly estimates pose and shape from depth. During testing, the estimated shape parameters were used to build up the dissimilarity matrix. BodyNet (BN) was applied in three different variants which are specified as follows:

1. BN1 (naive) Each shape vector is subtracted from those of the other samples and the sum of squared distances (SSD) is taken as a dissimilarity metric.
2. BN2 (augmented) The input at test-time were augmented by applying translations and horizontal flipping. This way, 50 shape vectors were estimated for each sample, which allowed to compute the centroid of the resulting parameter distribution. Intuitively, the centroid resembles a more robust shape descriptor than a single estimation. SSD was again used for computing the dissimilarities between the respective centroids.
3. BN3 (confidence weighted) After applying the test-time augmentation of BN2, the variance in the distribution of estimated parameters can be used to impose a confidence weighting. The variance vector v of each sample was normalized as $\hat{v} = \frac{v - \min(v)}{\max(v) - \min(v)}$ and $w = 1 - \hat{v}$ was used as weight vector. With element-wise multiplication ($\odot$), the dissimilarity computation for two samples s_1, s_2 changed to

$$fd_{1,2} = \frac{\sum (w_1 \odot w_2) \odot (\text{shapeVec}_1 - \text{shapeVec}_2)^2}{\sum w_1 \odot w_2}, \quad (3)$$

such that shape vectors were primarily compared at the parameters that were estimated with a higher confidence.

6. Experimental results

As a first result, given the dissimilarity matrices submitted, we compute the global retrieval scores, that are reported in Table 1

The first interesting result that can be seen is that lengths of skeletal branches are not informative at all. This can clearly be due to the large error in the segment fitting.

Retrieval scores are low, except for the NN, that, however, is certainly biased by the similarity of the meshes of the same subject with same clothing. To see this we also tested the average retrieval scores in the two subsets obtained with single illumination conditions. Results are shown in Table 1.

Figures 6 and 7 show the corresponding Precision-Recall plots, where it is clear that the recall values drop quickly when precision grows. DMS runs are consistently the best ones, showing that the combination of skeleton and shape holds a relevant amount of information even when the accuracy is low. The idea to estimate distances on skeletal points projected on the mesh might have reduced the error created by the inaccuracy of the skeleton estimation. DMS is also the only method tested that uses mesh information instead of point cloud data. This may also have had a role due to the removal of other noisy information from the original data.

	NN	FT	ST	E-m	DCG	mAp
LSS	0,01	0,022	0,046	0,031	0,314	0,056
SPC_Mean	0,693	0,165	0,224	0,152	0,505	0,164
SPC_STD	0,771	0,145	0,194	0,131	0,488	0,151
SPC_M+S	0,842	0,180	0,233	0,159	0,528	0,179
DMS-E	0,893	0,235	0,305	0,207	0,605	0,246
DMS-C	0,925	0,247	0,319	0,216	0,620	0,262
DMS-M	**0,931**	**0,249**	**0,325**	**0,220**	**0,624**	**0,267**
BN1	0,381	0,110	0,156	0,106	0,427	0,110
BN2	0,870	0,163	0,205	0,140	0,512	0,170
BN3	0,907	0,170	0,211	0,143	0,524	0,174

Table 1: *Retrieval scores obtained with submitted methods and baseline methods on the full models dataset.*

It may, in any case, appear strange that the more sophisticated approach (BN) does not provide the best results, but, actually the procedure of estimating shape parameters from simulated renderings may suffer of limits in pose sampling. The methods is in fact pose-biased as revealed by our further analyses. Biases and peculiarities of the different methods are indeed quite different and reveal interesting information.

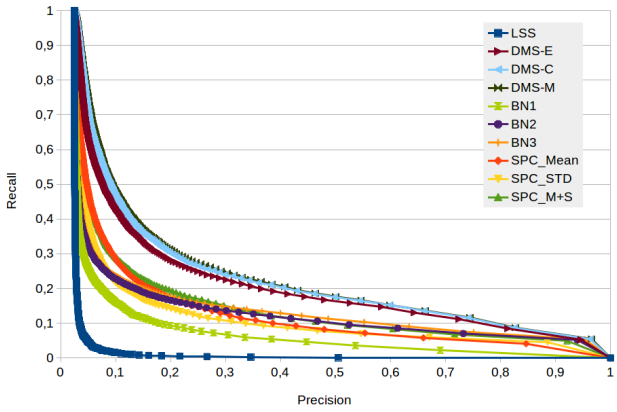

Figure 6: *Precision-recall plots for the retrieval with all the methods proposed on the full dataset.*

In order to understand this, we analyzed some figures derived from the best run of each group. If we consider, for example, the percentage of wrong first neighbor retrievals of each one (see Table 3), we see that the behavior of the algorithms is quite different: BN tends to retrieve more likely subjects in the same pose as the input while DMS tends to retrieve most likely subjects with same clothing.

	NN	FT	ST	E-m	DCG	mAp
LSS	0,026	0,025	0,047	0,036	0,268	0,059
SPC_Mean	0,172	0,094	0,150	0,092	0,408	0,088
SPC_STD	0,119	0,072	0,117	0,077	0,366	0,079
SPC_M+S	0,208	0,110	0,165	0,094	0,415	0,127
DMS-E	0,494	0,188	0,260	0,143	0,474	0,210
DMS-C	0,513	0,199	0,274	0,151	0,489	0,225
DMS-M	**0,521**	**0,204**	**0,279**	**0,155**	**0,494**	**0,231**
BN1	0,044	0,035	0,069	0,050	0,323	0,063
BN2	0,256	0,111	0,153	0,082	0,366	0,119
BN3	0,303	0,121	0,160	0,085	0,379	0,129

Table 2: *Averaged retrieval scores obtained with submitted methods and baseline methods on data subsets acquired with same illumination conditions*

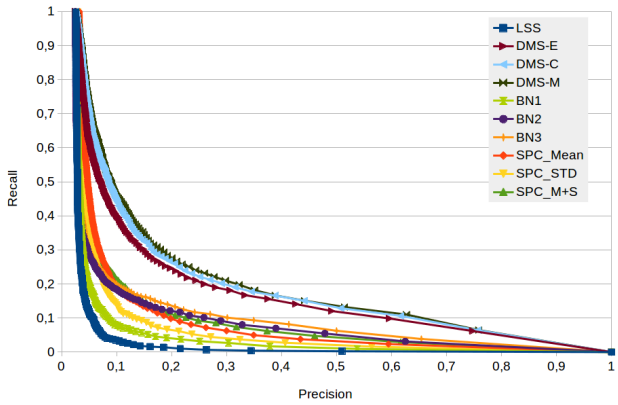

Figure 7: *Precision-recall plots for the retrieval with all the methods proposed on a subset of the original data with same illumination conditions*

If we consider the first 17 shapes retrieved, corresponding to the first tier statistics, and we look at the number of correctly labelled ones, we find other interesting insights. The number of retrieved shapes of the same subject is always far from the ideal one (17), but however, quite higher than the results that would have been obtained by random sampling. If we consider the number of retrieved models with the same pose and clothing and compare it with the expected number resulting in random sampling, of the input we see an evident bias, especially for pose, but different behaviors of different methods.

	SPC_M+S	BN3	DMS-M
same pose	76 %	95%	64 %
same clothing	55 %	39%	74 %
same lights	54 %	40%	64 %

Table 3: *Percentage of wrong NN retrievals sharing the same pose or the same clothing of the query model with the three best runs for each group.*

	SPC_M+S	BN3	DMS-M	Random exp.
same subject	3,06	2,88	4,27	0,40
same pose	11,85	15,10	12,09	5,67
same clothing	7,57	6,05	7,29	5,67
same light	8,06	8,05	8,06	8,47

Table 4: *Average number of retrieved models of same subject, same pose or same clothing among the first 17 shapes retrieved (1st tier), compared with the expected values in random retrieval.*

If we look at the correctly labelled retrieved models of this test, we see that all the methods are able to retrieve also examples of models of the same subject in different pose or with different clothes, with the exception of BN that seem not able to retrieve models in different poses. Table 5 shows the average number of correctly labelled retrieved shapes among the first 17 having different illumination, clothing and pose with respect to the query.

This fact tells us that, even if the task is quite hard, relevant information is encoded in the low quality scans and skeletons and probably the scores could be easily improved. One method of doing this could be simply merging the characterizations given by the different methods proposed here. This fact can be also deduced by looking at Figure 8. Here we plot confusion matrices obtained with classification of first (left), second (center) and third (right) retrieved models in the best runs of the different groups. Dark colors correspond to higher number of models retrieved, and dark spots not in the principal diagonals correspond to retrieved models not of the query subject. It is possible to see that the different methods often wrongly classify different subjects, and subjects that are mainly correctly retrieved by one method are often not well handled by others. This means that the algorithms exploit different, complementary body characteristics to identify subjects.

	SPC_M+S	BN3	DMS-M
diff clothing	1,61	1,86	2,75
different pose	0,95	0,08	1,22
different light	1,97	1,93	2,60
total correct	3,05	2,88	4,27

Table 5: *Average number of correct (same subject of the query) First-Tier retrieved models acquired in different poses, different clothing and different illumination conditions with respect to the query model.*

7. Discussion

The retrieval performances obtained in our task show that, as expected, it is hard to retrieve instance of the same human bodies from low resolution depth maps allowing change of pose and clothing and using only shape information without color.

All the methods proposed could be, however relevantly improved, and this is quite obvious, as the amount of time available for the test was quite small. Statistics on point clusters (SPC) can be enhanced adding more estimations. As the accuracy of the given

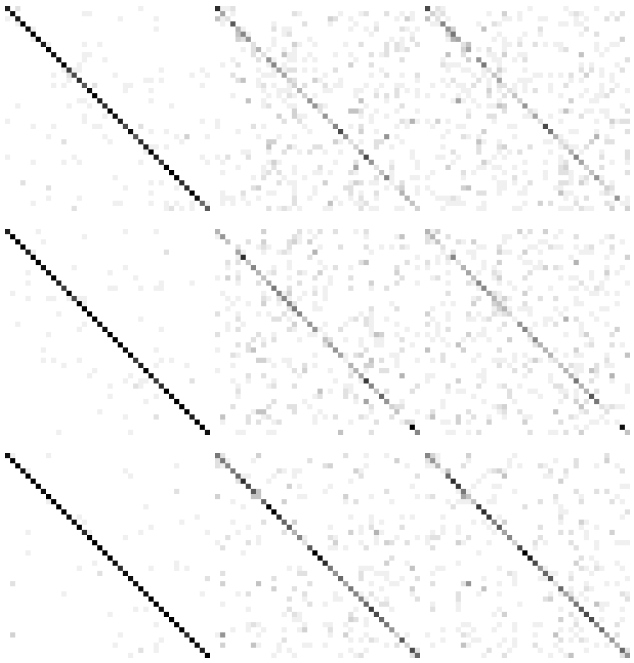

Figure 8: *Top row: confusion matrices obtained with classification of first (left), second (center) and third (right) retrieved model in the best SPC run. Middle row: confusion matrices obtained with classification of first (left), second (center) and third (right) retrieved model in the best BN run. Bottom row: confusion matrices obtained with classification of first (left), second (center) and third (right) retrieved model in the best DMS run.*

skeletal segments is low, this may be not used to estimate descriptors, but only for clustering, estimating as descriptors coefficients of cylinder or ellipsoid fitting for example. Furthermore, learning can be applied by training optimal feature space projection for classifying example data with known subject labelling. Similar considerations hold also for DMC descriptors.

BodyNet based methods could be improved in several ways, e.g. with a denser sampling of training poses. The huge effect of the augmentation step demonstrate the sensitivity of the estimated parameters on the input pose and the possibility of enhancing the results with simple heuristics.

Our analysis also showed that the different approaches proposed are in some sense complementary, each one performing better on different subjects and different conditions. This means that a smart feature fusion technique could also be successful in joining the different descriptors into a single one providing better scores.

Finally, supervised learning, that demonstrated a great effect in improving the retrieval performances of shape descriptors applied to whole human body scans [LBBC14] could surely be used also on to enhance the retrieval scores. As the dataset presents a sufficient number of models and a training set with different subjects with respect to the tested one is available, we think that interested researchers could surely test different supervised approach for this task with a relevant possibility of enhancing the retrieval scores.

In any case, we think that the insights coming from the analysis performed on the results can be extremely useful for the design of effective real-world applications.

It should be considered that the quality of the depth images is increasing and future generations of sensors and API may provide more precise reconstruction and improved algorithms could provide better skeleton estimates.

Furthermore, it should be considered that, in real world applications, color information could be used as well as dynamic information, here not exploited. It is clear that, using the time evolution of the point clouds, it is possible to enhance the quality of the human body characterization obtained with a single depth sensor. A recent work by Ichim et al. [IT16] showed, for example, the possibility of reconstructing accurate parametric reconstructions of human bodies from time evolving depth images, while in this context parametric reconstruction quality was certainly limited by the use of a single shot approach.

References

[BCDB*12] 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* (October 2012). 1

[BMR*99] BERNARDINI F., MITTLEMAN J., RUSHMEIER H., SILVA C., TAUBIN G.: The ball-pivoting algorithm for surface reconstruction. *Visualization and Computer Graphics, IEEE Transactions on 5*, 4 (1999), 349–359. 2

[BRM08] BASTIONI M., RE S., MISRA S.: Ideas and methods for modeling 3d human figures: the principal algorithms used by makehuman and their implementation in a new approach to parametric modeling. In *Proceedings of the 1st Bangalore Annual Compute Conference* (2008), ACM, p. 10. 4

[CCC*08] CIGNONI P., CALLIERI M., CORSINI M., DELLEPIANE M., GANOVELLI F., RANZUGLIA G.: Meshlab: an open-source mesh processing tool. In *Eurographics Italian Chapter Conference* (2008), vol. 2008, pp. 129–136. 2

[IT16] ICHIM A. E., TOMBARI F.: Semantic parametric body shape estimation from noisy depth sequences. *Robot. Auton. Syst. 75*, PB (Jan. 2016), 539–549. URL: http://dx.doi.org/10.1016/j.robot.2015.09.029, doi:10.1016/j.robot.2015.09.029. 6

[LBBC14] LITMAN R., BRONSTEIN A., BRONSTEIN M., CASTELLANI U.: Supervised learning of bag-of-features shape descriptors using sparse coding. In *Computer Graphics Forum* (2014), vol. 33, Wiley Online Library, pp. 127–136. 6

[MFB*14] MUNARO M., FOSSATI A., BASSO A., MENEGATTI E., VAN GOOL L.: One-shot person re-identification with a consumer depth camera. In *Person Re-Identification*. Springer, 2014, pp. 161–181. 1

[PSR*14] PICKUP D., SUN X., ROSIN P., MARTIN R., ET AL.: Shrec'14 track: Shape retrieval of non-rigid 3d human models. In *Proc. 3DOR* (2014), vol. 4, pp. 101–110. 1

[PSS99] PIGHIN F., SZELISKI R., SALESIN D. H.: Resynthesizing facial animation through 3d model-based tracking. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on* (1999), vol. 1, IEEE, pp. 143–150. 4

[SMKF04] SHILANE P., MIN P., KAZHDAN M., FUNKHOUSER T.: The princeton shape benchmark. In *Shape modeling applications, 2004. Proceedings* (2004), IEEE, pp. 167–178. 2

[WMKS*15] WANG J., MA K., KUMAR SINGH V., HUANG T., CHEN T.: Bodyprint: Pose invariant 3d shape matching of human bodies. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1591–1599. 1