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).
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 [CCC08] implementation of the ball pivoting algorithm [BMR99] (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.

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 Cumulated Gain (DCG).

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

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

Here is a detailed description of the proposed algorithm.

1. For each \(v \in \text{Ske}\) compute its closest point in the mesh, i.e.

\[
\text{Proj}_{\text{Mesh}}(v) = \arg \min_{w \in \text{Mesh}} ||v - w||_2^2
\]

2. For every pair of points in the skeleton compute the pairwise distance, i.e. \(D(i, j) = ||\text{Proj}_{\text{Mesh}}(v_i) - \text{Proj}_{\text{Mesh}}(v_j)||_2\) for all \(i, j = 1, \ldots, |\text{Ske}|\).
3. As \(D\) is symmetric, authors used \(\text{vec}(\text{Upper}(D))\) as descriptor, where \(\text{Upper}\) denotes the upper-triangular matrix capturing all the values above the diagonal and \(\text{vec}\) is the vectorization operator.

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

\[
\text{Dist}(a, b) = \frac{1}{|\text{Ske}|} \sum_{i=1, j=1}^{|\text{Ske}|} \left( ||\text{Upper}(D_a)(i, j) - \text{Upper}(D_b)(i, j)||_p \right)^{1/p}
\]

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 $\hat{v}$ of each sample was normalized as $\hat{v} = \frac{\max(v) - \min(v)}{\max(v) - \min(v)}$ 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

$$f_{d_{12}} = \frac{\sum (w_1 \odot w_2) \odot (\text{shapeVec}_1 - \text{shapeVec}_2)^2}{\sum w_1 \odot w_2},$$

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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>NN</th>
<th>FT</th>
<th>ST</th>
<th>E-m</th>
<th>DCG</th>
<th>mApx</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSS</td>
<td>0.01</td>
<td>0.022</td>
<td>0.046</td>
<td>0.031</td>
<td>0.314</td>
<td>0.056</td>
</tr>
<tr>
<td>SPC_Mean</td>
<td>0.693</td>
<td>0.165</td>
<td>0.224</td>
<td>0.152</td>
<td>0.505</td>
<td>0.164</td>
</tr>
<tr>
<td>SPC_STD</td>
<td>0.771</td>
<td>0.145</td>
<td>0.194</td>
<td>0.131</td>
<td>0.488</td>
<td>0.151</td>
</tr>
<tr>
<td>SPC_M+S</td>
<td>0.842</td>
<td>0.180</td>
<td>0.233</td>
<td>0.159</td>
<td>0.528</td>
<td>0.179</td>
</tr>
<tr>
<td>DMS-E</td>
<td>0.893</td>
<td>0.235</td>
<td>0.305</td>
<td>0.207</td>
<td>0.605</td>
<td>0.246</td>
</tr>
<tr>
<td>DMS-C</td>
<td>0.925</td>
<td>0.247</td>
<td>0.319</td>
<td>0.216</td>
<td>0.620</td>
<td>0.262</td>
</tr>
<tr>
<td>DMS-M</td>
<td><strong>0.931</strong></td>
<td><strong>0.249</strong></td>
<td><strong>0.325</strong></td>
<td><strong>0.220</strong></td>
<td><strong>0.624</strong></td>
<td><strong>0.267</strong></td>
</tr>
<tr>
<td>BN1</td>
<td>0.381</td>
<td>0.110</td>
<td>0.156</td>
<td>0.106</td>
<td>0.427</td>
<td>0.110</td>
</tr>
<tr>
<td>BN2</td>
<td>0.870</td>
<td>0.163</td>
<td>0.205</td>
<td>0.140</td>
<td>0.512</td>
<td>0.170</td>
</tr>
<tr>
<td>BN3</td>
<td>0.907</td>
<td>0.170</td>
<td>0.211</td>
<td>0.143</td>
<td>0.524</td>
<td>0.174</td>
</tr>
</tbody>
</table>

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

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

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.

<table>
<thead>
<tr>
<th></th>
<th>SPC_M+S</th>
<th>BN3</th>
<th>DMS-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>same pose</td>
<td>76 %</td>
<td>95%</td>
<td>64 %</td>
</tr>
<tr>
<td>same clothing</td>
<td>55 %</td>
<td>39%</td>
<td>74 %</td>
</tr>
<tr>
<td>same lights</td>
<td>54 %</td>
<td>40%</td>
<td>64 %</td>
</tr>
</tbody>
</table>

7. Discussion

The retrieval performances obtained in our task show that, as expected, it is hard to retrieve instances 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

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.

<table>
<thead>
<tr>
<th></th>
<th>SPC_M+S</th>
<th>BN3</th>
<th>DMS-M</th>
<th>Random exp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>same subject</td>
<td>3.06</td>
<td>2.88</td>
<td>4.27</td>
<td>0.40</td>
</tr>
<tr>
<td>same pose</td>
<td>11.85</td>
<td>15.10</td>
<td>12.09</td>
<td>5.67</td>
</tr>
<tr>
<td>same clothing</td>
<td>7.57</td>
<td>6.05</td>
<td>7.29</td>
<td>5.67</td>
</tr>
<tr>
<td>same light</td>
<td>8.06</td>
<td>8.05</td>
<td>8.06</td>
<td>8.47</td>
</tr>
</tbody>
</table>

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.
A.Giachetti et al. / Shrec’16 Track: Retrieval of Human Subjects from Depth Sensor Data

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


