In Proceedings of the 17th International Conference on Human-Computer Interaction, Los Angeles, CA, USA, August 2-7, 2015.

Interactive 3D digitization, retrieval, and analysis of ancient sculptures, using infrared depth sensors for mobile devices

Angelos Barmpoutis[1]*, Eleni Bozia[1], and Daniele Fortuna[2]

[1] University of Florida, Gainesville FL 32611, USA,
angelos@digitalworlds.ufl.edu, bozia@ufl.edu
[2] Museo Nazionale Romano di Palazzo Altemps, Roma 00186, Italia,
daniele.fortuna@beniculturali.it

Abstract. In this paper a novel framework is presented for interactive feature-based retrieval and visualization of human statues, using depth sensors for mobile devices. A skeletal model is fitted to the depth image of a statue or human body in general and is used as a feature vector that captures the pose variations in a given collection of skeleton data. A scale- and twist- invariant distance function is defined in the feature space and is employed in a topology-preserving low-dimensional lattice mapping framework. The user can interact with this self-organizing map by submitting queries in the form of a skeleton from a statue or a human body. The proposed methods are demonstrated in a real dataset of 3D digitized Graeco-Roman statues from Palazzo Altemps.

Keywords: Depth sensors, RGB-D, Kinect, 3D Object Retrieval, Digital Humanities, Statues, Museum Studies

1 Introduction

For the past decade, the technological advances in the areas of portable electronic devices have revolutionized the use and range of applications of tablet computers, smart phones, and wearable devices. Furthermore, various multimodal sensors have been introduced in mobile devices to offer more natural user-machine interactions. Depth sensors (range cameras) have become popular as natural user interfaces for desktop computers and have recently become available for mobile devices, such as the Structure Sensor^{TM}by Occipital.

Depth sensors have been used in various applications related to body tracking such as human detection, model-based 3D tracking of hand articulations [11], human pose recognition and tracking of body parts [12], real-time 3D reconstruction of the articulated human body [3], motion tracking for physical therapy [5],

* This project was in part funded by the Rothman Fellowship from the Center for the Humanities and the Public Sphere and the research incentive award from the College of the Arts at the University of Florida.

and others [7]. The reader is referred to [7] for a more detailed review of RGB-D applications.

In this paper we present a novel application of depth sensors for mobile devices in the topic of digital archaeology. Digital technologies have been adopted in various areas related to museum experience, digital preservation, as well as digitization and study of archaeological artifacts [4].

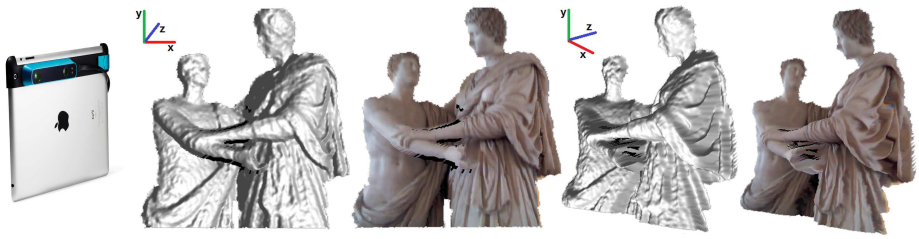

Fig. 1. Example of an RGB-D frame captured by the camera and depth sensor for a tablet computer (shown on the left). The same depth frame is rendered from two different perspectives —with and without color texture. Images shown with permission from the Italian Ministry of heritage, cultural activities and tourism. Su concessione del Ministero dei beni e delle attività culturali e del turismo - Soprintendenza Speciale per il Colosseo, il Museo Nazionale Romano e l'area archeologica di Roma.

Digital collections become even more useful educationally and scientifically when they provide tools for searching through the collection and analyzing, comparing, and studying their records. For example an image collection becomes powerful if it can be searched by content, technique, pattern, color, or even similarity with a sample image. The lack of keywords and generalizable annotation for such type of analysis generates the need for keyword-free feature-based analysis.

In this paper a framework is presented for 3D digitization, database retrieval, and analysis of classical statues using depth sensors for mobile devices. In this framework each statue is represented in a feature space based on the skeletal geometry of the human body. A distance function is defined in the feature space and is employed in order to find statues with similarities in their pose. The search query in the presented framework is the body of the user, who can interact with the system and find which statues have poses similar to the user's pose. The proposed methods are demonstrated, using real data from classical statues (shown in Fig. 1) collected in Palazzo Altemps in Rome, Italy.

The contributions in this paper are threefold: a) A novel application of depth sensors for mobile devices is presented for feature-based retrieval of 3D digitized statues. b) A scale- and twist-invariant distance function between two given skeletons is proposed. c) A special type of self-organized maps is presented for interactive visualization of the space of body postures in a low-dimensional lattice.

2 Methods

Depth sensors can be used to detect the presence of a particular skeletal geometry, such as human skeletal geometry, by fitting to each acquired depth frame a skeletal model that consists of the following set of parameters:

$$\mathcal{S} = \{\mathbf{t}_j \in \mathbb{R}^3, \mathbf{R}_j \in SO(3) : j \in \mathcal{J}\} \tag{1}$$

where $\mathcal{J}$ is a set of indices of joints connected together in a tree structure of parent/children nodes. Each joint is defined by its location in the 3D space, which is expressed as a translation $\mathbf{t}_j$ from the origin of the coordinate system of the root joint, and its orientation in the 3D space is given as rotation matrix $\mathbf{R}_j$ with respect to the orientation of the root node. There are several algorithms that compute $\mathcal{S}$ from RGB-D, such as those implemented in the Microsoft Kinect SDK [1], in OpenNI library [2], and others [14, 12].

2.1 Skeleton distance functions

One way to compute distances between unit vectors is the so-called *cosine distance* given by $1-\cos(\phi)$, where ϕ is the angle between the two vectors. Although the triangle inequality property is not satisfied by this function and therefore is not considered a distance metric, it is computationally very efficient as it can be expressed in a polynomial form. Cosine distance can be extended in order to perform comparisons between elements of $SO(n)$ space by calculating the cosine of the angles between the corresponding rotated orthogonal basis as follows:

$$dist(\mathbf{R}_1, \mathbf{R}_2) = 3 - \cos(\phi_1) - \cos(\phi_2) - \cos(\phi_3) = 3 - trace({\mathbf{R}_1}^T \mathbf{R}_2) \tag{2}$$

where ϕ_i denotes the angle between the rotated basis vectors $\mathbf{R}_1\mathbf{e}_i$ and $\mathbf{R}_2\mathbf{e}_i$. It can be easily shown that the value of Eq. 2 becomes zero when $\mathbf{R}_1 = \mathbf{R}_2$.

In the case of skeletal geometry, the distance between two poses a and $b \in \mathcal{S}$ can be computed by evaluating Eq. 2 for every joint in $\mathcal{J}$. Such a distance function is scale invariant since it does not take under consideration the locations of the joints, which is a desirable property for our application. Furthermore, the calculated distance can become twist invariant (i.e. invariant under rotations around the line segment that connects two joints) by evaluating Eq. 2 only for the basis vector that corresponds to the axis along the particular line segment as follows:

$$dist(a, b) = |\mathcal{J}| - \mathbf{e}_1^T \sum_{j \in \mathcal{J}} {\mathbf{R}_j^a}^T \mathbf{R}_j^b \mathbf{e}_1 \tag{3}$$

where $|\mathcal{J}|$ is the cardinality of the set of tracked joints, and $\mathbf{R}_j^a$, $\mathbf{R}_j^b$ are the corresponding rotation matrices of the skeletons a and b respectively. Without loss of generality, $\mathbf{e}_1$ is a unit vector that denotes the basis of a line segment in the skeletal structure. Eq. 3 is scale-invariant and twist-invariant, which are both necessary properties in our application. Scale-invariance guarantees that the distance between skeletons of different subjects will be zero if both are in

the same pose. Twist-invariance makes the function robust to possible miscalculations of the rotation of each joint during the skeleton fitting process. In the next section we employ Eq. 3 to achieve 3D object retrieval from a database of human statues using self-organizing maps.

2.2 Interactive statue retrieval using self-organizing skeletal maps

Given a dataset of skeletons s_1, s_2, $\cdots \in \mathcal{S}$ and a query skeleton $q \in \mathcal{S}$, we need to construct a topographic mapping to a 2-dimensional lattice that satisfies the following 2 conditions: a) q is mapped to a fixed location at the center of the lattice, and b) similar skeletons should be mapped to neighboring lattice locations. The goal of such mapping is to generate an interactive low-dimensional visualization of the multi-dimensional manifold $\mathcal{S}$ for 3D statue retrieval purposes. The user can provide an input query q, which could either belong to the existing dataset s_i (i.e. the posture of a previously digitized statue) or be a new sample in the space $\mathcal{S}$ (i.e. the posture of a human subject or a new statue).

Self-organizing maps have been well-studied in literature and have been employed in various applications related to machine learning and data visualization [9, 6, 8, 13]. A self-organizing map is a type of artificial neural network originally proposed by T. Kohonen [9] that consists of a set of nodes, each of which is located at position x of a low-dimensional lattice (in our application $x \in \mathbb{R}^2$) and is associated with an unknown weight vector in the original feature space (in our application $w_x \in \mathcal{S}$).

The weight vectors produce a dynamic mapping $\mathbb{R}^2 \rightarrow \mathcal{S}$ in the form $f(x) = \sum_y w_y K(|x-y|)$, where K is a neighborhood-based kernel function centered at $y \in \mathbb{R}^2$. The mapping is modified by following an iterative energy optimization process using the following update rule:

$$w'_x = w_x - \alpha(t) \frac{\partial dist(s_i, w_x)}{\partial w_x} K(|x - x^*|) \tag{4}$$

where $x^* = \arg\min_x dist(s_i, w_x)$. The derivative of the distance function defined in Eq. 3 can be analytically calculated and results in a $|\mathcal{J}| \times 3$-size gradient vector that contains the coefficients of the vectors $-\mathbf{R}^i_j \mathbf{e}_1 \ \forall j \in \mathcal{J}$, where $\mathbf{R}^i_j$ are the rotation matrices of $s_i \in \mathcal{S}$. Therefore, in our implementation the weight space is $|\mathcal{J}| \times 3$-dimensional and consists of $|\mathcal{J}|$ unit vectors. This mapping of the feature space is due to the scale- and twist-invariance of the distance function, as discussed previously.

Eq. 4 is applied iteratively for all s_i in the given dataset and for all lattice locations x except for a predefined central node that corresponds to the query skeleton q (i.e. $w_x = q$). The rest of the weights can be initialized randomly using a Gaussian distribution centered at q. After each iteration the w_x is properly normalized in order to ensure that the components of the weight vector correspond to $|\mathcal{J}|$ unit vectors.

In the next section we demonstrate the presented techniques, using a real dataset of digitized sculptures.

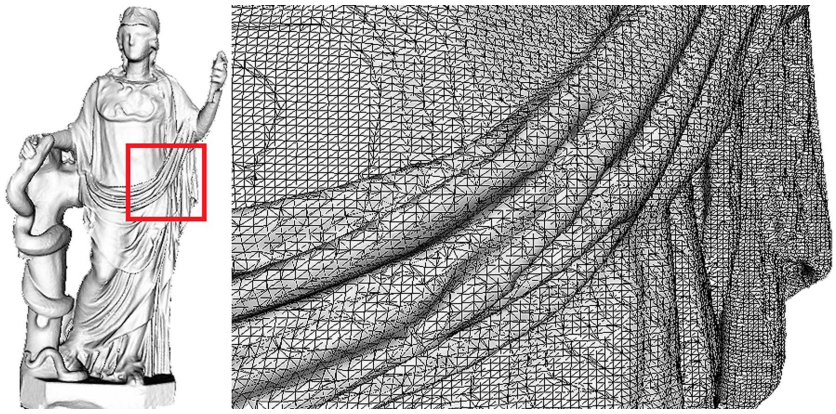

Fig. 2. Left: Example of a statue reconstructed in 3D by fusing a sequence of depth frames. Right: Zoomed view of the reconstructed mesh to show detail.

3 Experimental Results

In this project we used the Structure Sensor^{TM}by Occipital, which was attached in front of a tablet computer (iPad Air^{TM}by Apple). The resolution of the depth sensor was 640×480 pixels at 30 frames per second and was calibrated so that it records depth in the range from 0.4m to 3.0m, which is adequate for capturing life-size statues. Another depth sensor, Kinect^{TM}by Microsoft, was also used in our depth fusion experiments, which were performed on a 64-bit computer with Intel Core i7^{TM}CPU at 2.80GHz and 8GB RAM. Both Kinect and Structure sensors had similar resolution, range of operation, and field of view and were seamlessly used in this project, and therefore we will not differentiate their depth data in our discussion.

In order to create a test dataset for our experiments we digitized in 3D statues from the collection of Palazzo Altemps in Rome, Italy with the permission of the director of the museum, Alessandra Capodiferro. Palazzo Altemps is located in the centre of Renaissance Rome, between Piazza Navona and the Tiber river, in the northern part of Campus Martius. Archaeological excavations have uncovered Roman structures and finds dating from the 1st century AD to the modern age. The current building remained property of the Altemps family for about three centuries, after it was originally acquired by Cardinal Marcus Sitticus Altemps in 1568, who commissioned architects and artists of the time to undertake significant work to extend and decorate the palace. Today the National Roman Museum branch at Palazzo Altemps houses important collections of antiquities, consisting of Greek and Roman sculptures that belonged to various families of the Roman aristocracy in the 16th and 17th centuries [10].

This study focused on statues from three of the collections housed in the museum: Boncompagni Ludovisi collection, Mattei collection, and Altemps collection. Boncompagni Ludovisi is the famous 17th century collection of ancient

Fig. 3. This figure shows selected samples from our dataset of 3D digitized statues.

Fig. 4. Visualization of the dataset of skeletons shown from three different perspectives.

sculptures housed in Villa Ludovisi in Quirinal , which was a popular attraction for scholars, artists, and travelers from all over the world up to and throughout 19th century. The 16th century Villa Celimontana with its Navicella garden was property of Ciriaco Mattei and was decorated with ancient sculptures, some of which are today in the Mattei collection of Palazzo Altemps. Finally, the important antiquities collection of the Altemps family consists of about a hundred noteworthy pieces of sculptures. These decorated the aristocratic home of Marcus Sitticus Altemps (the grandson cardinal of Pope Pius IV) in line with the prevailing antiquarian taste of the 16th century.

The statues of our interest were in various standing positions and were portraying figures from classical mythology. In total 22 life-size statues were digitized by manually moving the depth sensor around each statue and fusing the acquired data using the Kinect Fusion software [1]. The 3D digitization process lasted for approximately 3 minutes for each statue ($181.5 \pm 47.2 sec$). In addition to the 3D reconstructed models, a dataset of the corresponding skeletons was also created by using the skeleton fitting process of the OpenNI library [2] with auto-calibration setting. In the case of fitting errors caused by the presence of adjacent objects or heavy clothing the estimated skeletons were manually corrected. The dataset created from this process is shown in Figs. 3 and 4.

Fig. 4 shows the skeletal samples $s_1, \ldots, s_{22} \in \mathcal{S}$ in the dataset. The skeletons in this plot were normalized in terms of their sizes (full body size and limb size) in order to show the variability of the poses in the dataset. Based on the plots from all three perspectives, it is evident that larger differences are observed in the position of the arms of the statues and smaller yet notable variations are observed in the orientations of the legs and torso as expected.

In the proposed framework, the pose of the statues forms the feature space, which is employed in the comparisons between statues and feature-based database searches for statues with similar poses. The quality of the employed features is determined by their ability to capture the distinct characteristics of each element in the search space; hence the pose variations in Fig. 4 are essential in the proposed framework.

In order to test the interactive statue retrieval framework presented in Sec. 2 we performed skeleton fitting to a human subject who stood in a particular pose in front of the depth sensor. The fitted skeleton was provided as the query input q to a self-organizing 2D map of size 5×3. The central node in the map was assigned to the query vector. The rest of the map was initialized randomly and updated for 1000 iterations, a process that was completed in less than 1 sec in the tablet computer. A representative result is shown in Fig. 5.

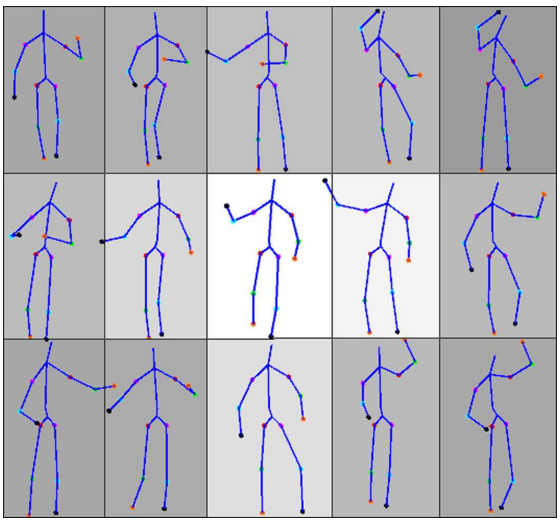

Fig. 5. An example of a self-organized map of skeletons. The map was generated around the query skeleton located at the center of the map. The background intensity shows the distance from the query skeleton.

By observing Fig. 5 we can see that skeletons with similar poses were mapped in adjacent locations on the map, such as the adjacent skeletons in the center of the map (see also the skeletons in the upper right and the lower right corners).

Furthermore, there are smooth transitions between different poses when possible. At this point it should be noted that due to the limited number of samples in our dataset it is not always possible to put the samples in a smooth order.

The gray-scale intensity of the background in Fig. 5 corresponds to value of the distance (given by Eq. 3) between each skeleton and the query skeleton. As expected, the locations around the center of the map correspond to smaller distances (brighter intensities) compared to the areas along the edges of the map.

Finally, Fig. 6 shows the three "closest" statues to the given skeleton query among all statues in the database. The corresponding distance values are also reported. Although the two first statues have similar postures on the upper part of the body, the pose of the first statue's legs better resembles the query skeleton. As a result, the first statue has the smallest distance from the query, which demonstrates the efficacy of the presented framework.

Fig. 6. Demonstration of interactive search. The search query is shown on the left. The best matching statues and their corresponding distance from the query are reported.

4 Conclusion

The pilot study in this paper shows that the presented framework can be used for keyword-free feature-based retrieval of statues in mobile devices. This has the potential to be used as an interactive guide in museums, but also as a scientific tool that assists scholars in identifying statues with similar characteristics from a large repository of statues. The future use of depth sensors in mobile devices will significantly support the creation of such repositories of 3D digitized artifacts, using limited resources (in terms of scanning time, computational effort, and cost) as well as their computer-assisted study as demonstrated in this paper.

Acknowledgement. The authors would like to acknowledge Alessandra Capodiferro for providing permission to perform this study in Palazzo

Altemps and the Italian Ministry of heritage, cultural activities and tourism for providing permission to publish in this paper the collected data. This project would not be possible without the funding support by the Rothman Fellowship in the Humanities to Eleni Bozia from the Center for the Humanities and the Public Sphere at the University of Florida and the research incentive award to Angelos Barmpoutis from the College of the Arts at the University of Florida. The authors would like to thank the sponsors and the anonymous reviewers who provided insightful comments and suggestions.

References

1. Microsoft Kinect SDK. http://www.microsoft.com/en-us/kinectforwindows/
2. OpenNI. http://www.openni.org/
3. Barmpoutis, A.: Tensor body: Real-time reconstruction of the human body and avatar synthesis from RGB-D. IEEE Transactions on Cybernetics 43(5), 1347–1356 (2013)
4. Barmpoutis, A., Bozia, E., Wagman, R.S.: A novel framework for 3d reconstruction and analysis of ancient inscriptions. Journal of Machine Vision and Applications 21(6), 989–998 (2010)
5. Barmpoutis, A., Fox, E., Elsner, I., Flynn, S.: Augmented-reality environment for locomotor training in children with neurological injuries. In LNCS 8678 (Springer) Proceedings of the Workshop on Augmented Environments for Computed Assisted Interventions pp. 108–117 (14 September 2014)
6. Bishop, C.M., Svensen, M., Williams, C.K.I.: The generative topographic mapping. Neural Computation 10(1), 215–234 (1998)
7. Han, J., at al.: Enhanced computer vision with Microsoft Kinect sensor: A review. IEEE Transactions on Cybernetics 43(5), 1318 – 1334 (2013)
8. Haykin, S.: Neural Networks and Learning Machines. Prentice Hall, 3 edn. (2008)
9. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43(1), 5969 (1982)
10. La Regina, A.: Museo Nationale Romano. Soprintendenza Archaeologica di Roma. Mondadori Electa S.p.A. Milan (2005)
11. Oikonomidis, I., et al.: Efficient model-based 3d tracking of hand articulations using Kinect. In Proc. of the British Machine Vision Association Conference (2011)
12. Shotton, J., et al.: Real-time human pose recognition in parts from single depth images. In: IEEE CVPR Conference. pp. 1297–1304 (2011)
13. Ultsch, A.: Emergence in self-organizing feature maps. In Ritter, H.; Haschke, R. Proceedings of the 6th International Workshop on Self-Organizing Maps (2007)
14. Xia, L., et al.: Human detection using depth information by Kinect. IEEE Conference on Computer Vision and Pattern Recognition Workshops pp. 15–22 (2011)