Automated Human Avatar Synthesis for Obesity Control using Low-Cost Depth Cameras

Angelos BARMPOUTIS[a,1]
[a] *Digital Worlds Institute, University of Florida, angelos@digitalworlds.ufl.edu*

Abstract. In this paper a framework is presented for monitoring shape changes on the human body with applications to obesity control. This framework uses a low-cost infrared depth camera in order to capture the 3D shape of the human body and approximate it as a set of spherical functions.

Keywords. Tensor Basis, Avatar Synthesis, 3D reconstruction, Kinect

Introduction

Infrared depth cameras have been widely used as low-cost peripheral devices for various applications related to virtual reality interaction using natural user interfacing. The depth information captured on a daily basis by these devices can also be used to extract useful information related to the behavior and physical shape of the users, which can be associated with parameters related to the user's health and physical condition.

There are several examples in literature that present medical applications of depth cameras. A game-based rehabilitation system was presented in [9] using body tracking. A similar application was applied to kids with spinal cord injuries as a mechanism for exciting young patients to perform walking exercises [2]. Other medical applications of depth cameras include controller-free exploration of medical image data [8] for avoiding the spreading of germs caused by interacting with physical controller devices.

The aforementioned medical applications as well as the work presented in this paper employ several well studied principles from three-dimensional computer vision [5,6,7] in novel frameworks that provide the medical community with useful tools for rehabilitation, medical data navigation using natural user interfaces, and patient monitoring.

In this paper a framework is presented for estimating descriptive parameters of the human body shape using a low-cost depth camera in contrast to the traditional computer vision algorithms that attempt to reconstruct human avatars using image- or video-based approaches [12,13,14,15,16]. The proposed framework can be used for synthesizing human avatars as a set of spherical functions that approximate the shape of independent regions in the human body such as the torso, the head, the arms and the

[1] Corresponding Author. University of Florida, P.O. Box 115810, 101 Norman Gym, Gainesville, FL 32611-5810, USA. E-mail: angelos@digitalworlds.ufl.edu

legs. In the experimental section a novel application is presented for obesity monitoring over a period of time and quantitative comparison with the average human shape computed using the 3D spherical functions estimated by the proposed framework. One of the major goals of this project is to use a traditional gaming interface (such as depth cameras) in order to raise awareness of the importance of child obesity, which has reached 17% of the population aged 2-19 years old in the US [1]. Besides obesity control, the proposed framework can be used to monitor post-surgical changes in the shape of human body, during chemotherapy, as well as other types of medical treatment.

1. Methods & Materials

In this section a framework is presented for approximating the 3D shape of the human body with a set of homogeneous polynomial functions estimated using a depth and a video camera. A diagram of the proposed framework is shown in Fig. 1.

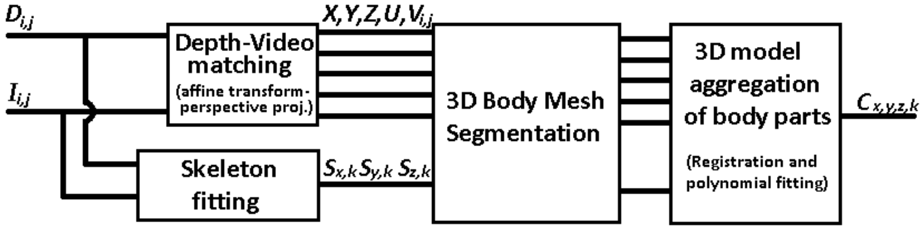
$D_{i,j}$
$I_{i,j}$
Depth-Video matching
(affine transform-perspective proj.)
$X,Y,Z,U,V_{i,j}$
Skeleton fitting
$S_{x,k} S_{y,k} S_{z,k}$
3D Body Mesh Segmentation
3D model aggregation of body parts
(Registration and polynomial fitting)
$C_{x,y,z,k}$

Figure 1. Block diagram of the proposed framework. The output of the overall system is a tensor coefficient vector that approximates the shape of independent regions in the human body.

The depth camera generates a sequence of discrete depth frames in the form of 2D arrays $\mathbf{D}_{ij}$, which can be equivalently expressed as a quad mesh given by $\mathbf{X}_{ij}=(i-i_c)\mathbf{D}_{ij}c_d^{-1}$, $\mathbf{Y}_{ij}=(j-j_c)\mathbf{D}_{ij}c_d^{-1}$, and $\mathbf{Z}_{ij}=\mathbf{D}_{ij}$, where i_c, j_c denote the coordinates of the central pixel in the depth frame, and c_d is the focal length of the depth camera.

The video frames can be associated with the 3D quad mesh by using texture mapping given by the coordinates $\mathbf{U}_{ij}=\mathbf{X'}_{ij}\mathbf{Z'}_{ij}^{-1}c_v$, $\mathbf{V}_{ij}=\mathbf{Y'}_{ij}\mathbf{Z'}_{ij}^{-1}c_v$, where the coordinates of the vector [X' Y' Z'] are related to [X Y Z] via a known rigid transformation (rotation and translation), and c_v is the focal length of the video camera. The aforementioned transformation corresponds to the mapping between the location and orientation of the two cameras.

The depth and video frames are provided as input to an algorithm that fits a simple human skeletal model representing the person depicted in the data [10]. The model consists of points in R^3 corresponding to the location of major joints in the human body, 13 of which are used by the proposed framework for segmenting the 3D quad mesh of each frame. The computed list of points $\mathbf{s}_k \in R^3$, k=1…13, forms 13 line segments depicted in Fig. 2 (fourth panel), 4 in the torso, 2 in each arm, 2 in each leg, and 1 in the head.

For every vertex $\boldsymbol{p}=[\mathbf{X}_{ij}\ \mathbf{Y}_{ij}\ \mathbf{D}_{ij}]$ in the quad mesh we compute its distance from each of the 13 line segments of the skeleton model as follows:

$$dist(\boldsymbol{p},\boldsymbol{a},\boldsymbol{b})=||\boldsymbol{a}+x(\boldsymbol{b}-\boldsymbol{a})-\boldsymbol{p}|| \qquad (1)$$

where $\boldsymbol{a},\boldsymbol{b} \in R^3$ are vertices/joints that define a particular line segment in the skeleton model, and x is the projection of $\boldsymbol{p}$ onto the line segment given by:

$$x=max\{min\{(\boldsymbol{b}-\boldsymbol{a})\cdot(\boldsymbol{p}-\boldsymbol{a})/||\boldsymbol{b}-\boldsymbol{a}||^2,1\},0\} \qquad (2)$$

The *max* and *min* functions in Eq. 2 guarantee that if the projection falls outside the line segment, the distance computed by Eq. 1 will be equal to the Euclidean distance between $\boldsymbol{p}$ and the closest end-point of the line segment (i.e $min\{||\boldsymbol{a}-\boldsymbol{p}||,||\boldsymbol{b}-\boldsymbol{p}||\}$). Using the distance measure defined by Eq. 1 every vertex p is assigned to the closest body region. The quad mesh segmentation is performed for every frame as demonstrated in Fig. 2. Note that the points that do not belong to the depicted human subject can be easily thresholded across Z_{ij}, since the background objects usually have larger D_{ij} values.

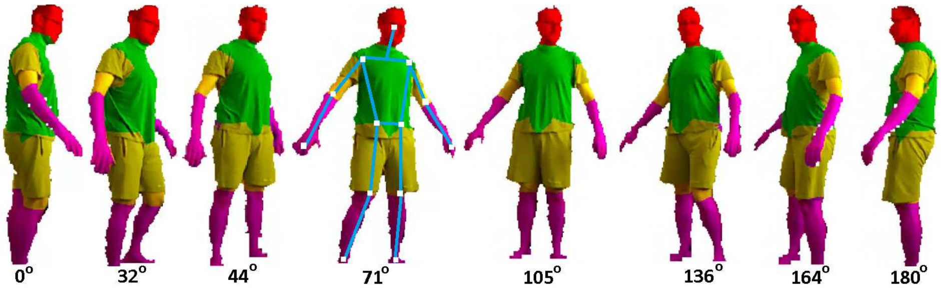

Figure 2. Segmentation examples of the 3D quad meshes recorded during a 360 degree rotation of the depicted human subject. The fitted 13 point skeleton is shown on the fourth example from the left.

After the segmentation step, a time sequence of quad meshes for each body region was generated. The meshes in each of the obtained sequences can be considered transformed versions of the two neighboring in the time domain meshes. Assuming that the individual body regions are not significantly deformed in two consecutive time frames, the transformation can be modeled as a rigid motion 4x4 matrix with 6 degrees of freedom (close to the identity matrix), which can be easily estimated using a distance measure between the corresponding point sets [11] as well as the intensity maps of the corresponding textures. The registration process can be performed in parallel for all the neighboring meshes in order to compute a globally registered point set for each body region.

In total 10 point sets will be constructed: two for each arm, two for each leg, one for the torso, and one for the head. Each of the point sets can be approximated by any continuous spherical function that can be modeled using spherical harmonic bases or their equivalent Cartesian tensor basis [3] as follows:

$$f_k(\boldsymbol{v})=\sum_{x,y,z}\boldsymbol{C}_{xyzk}\, v_1^x\, v_2^y\, v_3^z \qquad (3)$$

where $k=1\ldots10$ is the index of the body region, $\boldsymbol{C}_{xyzk}$ is the vector of unknown coefficients and $\boldsymbol{v}=[v_1\ v_2\ v_3]^T$ is a unit vector. The summation in Eq. 3 is over the powers x,y,z of the tensor bases that define the order of the approximation. In order to approximate any arbitrary function, one even and one odd order should be included. In the experiments presented in next section the summation in Eq. 3 was implemented for all non-negative x,y,z that satisfy $x+y+z=10$ and $x+y+z=9$. The unknown coefficients $\boldsymbol{C}_{xyzk}$ can be estimated by fitting Eq. 3 to the points $\boldsymbol{p}_i$ of each point set by minimizing [4]:

$$E(\boldsymbol{C}_{xyzk})=\sum_i \Big(||\boldsymbol{p}_i-\boldsymbol{\mu}||-f_k(\boldsymbol{p}_i-\boldsymbol{\mu}/||\boldsymbol{p}_i-\boldsymbol{\mu}||)\Big)^2 \qquad (4)$$

where $\boldsymbol{\mu}$ is the mean of the point set. The minimization of Eq.4 can be implemented as the least-squares solution to the linear system **Ac**=**b** where $\mathbf{b}_{\mathrm{i}}=||\boldsymbol{p}_i-\boldsymbol{\mu}||$ and $\mathbf{A}_{\mathrm{ij}}= v_1^x\ v_2^y\ v_3^z$ where $\boldsymbol{v}=\boldsymbol{p}_i-\boldsymbol{\mu}/||\boldsymbol{p}_i-\boldsymbol{\mu}||$. The fitted polynomial can be visualized as a spherical function demonstrated in Fig. 5. The proposed framework was applied to real data and the experimental results are discussed in the next section.

2. Results

In our experiments we used the PrimeSense infrared depth camera as well as the video camera of Microsoft's Kinect gaming control device that was connected with a computer via a USB 2.0 port. The resolution of the depth camera was 320x240 pixels with a viewing range from 0.8m to 4.0m and horizontal field-of-view angle of 57 degrees.

Based on the viewing volume defined by the above specifications, an average-height human body (~1.75m) when viewed entirely by this particular depth camera using the full vertical range of 240 pixels, it can be recorded with a 3D point accuracy limit of ~73mm+ε, where ε is an additional accuracy error term related to the signal-to-noise ratio of the sensor. In practice, due to the term ε as well as the fact that a moving subject cannot fully utilize the vertical field-of-view, the 3D point accuracy error is estimated at ~1cm, which is acceptable for the purpose of our proposed application. If one is interested in performing the proposed analysis to part of the human body only, for instance the upper part or the torso, the accuracy error can be reduced to ~0.43cm.

Five volunteers participated in our preliminary experiments. In order to focus our study to a specific age/gender group, all subjects were males, between 20-31 years old. The subjects were positioned in front of the depth camera in such a distance so that their full body was in the viewing volume. Then, the subjects were asked to rotate slowly around their position having their hands slightly raised in order to avoid biasing our results due to possible mis-segmentation between the torso and the hands.

The depth and video streams were recorded during each session, approximately 10 sec. stream for a 360 degree rotation, which corresponded to ~250-300 frames (i.e. 25-30 frames/second). At the same time, a simple skeleton model consisting of 20 joints was fitted to the data in real time using the algorithm provided by Microsoft's Kinect SDK and the computed skeleton stream was recorded along with the video and depth streams.

After finishing the data collection, each dataset was processed by the proposed framework as presented in the previous section and illustrated in Fig. 1. The data of each frame were converted to the form of a textured rectangular mesh ($X_{ij},Y_{ij},Z_{ij},U_{ij},V_{ij}$, $i=1\ldots320,\ j=1\ldots240$), which were segmented into 11 regions including to 10 body parts and the background. An example of the result of the body segmentation is shown in Fig. 2. The segmented 3D meshes are shown in different colors for various frames during the rotation of the subject. The corresponding segmented meshes from each frame were registered to each other producing a point cloud for each body segment, visualized in Fig. 3 as a full 3D body. Finally the point clouds can be approximated by a least-squares-fitted continuous spherical function using homogeneous polynomial basis. The computed polynomial coefficients can be used as a compact descriptor of the approximated 3D shapes for statistical analysis of a collection of similar shapes from a population of subjects as well as shapes computed from the same subject during a specific time period.

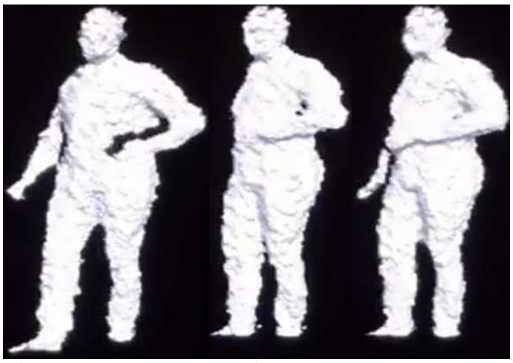

Figure 3. Visualization of the point set constructed after registering the quad meshes from adjacent frames.

In our preliminary experiments we computed the Euclidean average of the polynomial coefficient vectors from the torso of the five subjects. An example of visualizing the distance of a specific subject from the computed average is shown in Fig. 4. The locations that correspond to largest distance values are shown in red using the color map on the left of the same figure.

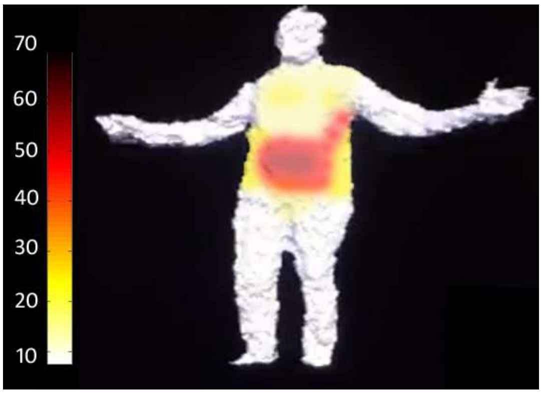

Figure 4. Example of visualizing the distance of a specific subject from the average torso shape. The regions highlighted in red correspond to larger differences.

Finally, the fitted polynomials as well as the distance between two polynomials can be visualized as spherical functions. In both cases, the value of the function can be computed for a predefined set of unit vectors uniformly distributed on the unit sphere. Such a set of unit vectors can be generated as the Nth-order tessellation of the icosahedrons on the unit sphere, which produces a triangular mesh that approximates the unit sphere (as shown in Fig. 5 left). After evaluating the spherical function of interest for all vertices on the triangular mesh, the function can be plotted by multiplying the magnitude of each vertex with the corresponding value of the function. Figure 5 (right) shows an example of visualizing the distance of the polynomial fitted to the torso of a specific subject from the average polynomial.

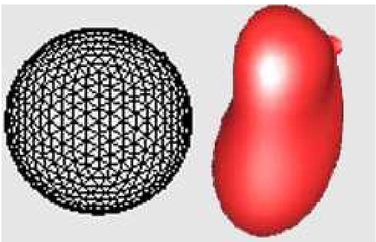

Figure 5. Plot of the estimated homogeneous polynomial as a 3D closed surface (right) using the triangular mesh on the left.

3. Conclusions & Discussion

The experimental results presented in the previous section demonstrate the efficacy of the proposed framework. The interactive nature of depth cameras as a tool for natural user interfaces in addition to their low cost and their popularity in the digital entertainment industry brings to the average consumer a device for various home-based medical applications, some of which discussed in the introductory section. The monitoring of the changes in our body and its comprehensive comparison with the average body shape of the corresponding age/gender group can be proven to be a significant tool against obesity or other related diseases such as heart disease.

In the future, we plan to apply the proposed framework to large datasets collected from various populations with critical body changes such as women during and after pregnancy as well as patients during various stages of chemotherapy. Furthermore, we plan to compute body shape atlases from healthy subjects of various ages, genders and ethnicities. Such an atlas could be used for analyzing quantitatively the shape differences of the body across population groups and derive useful statistical results.

References

[1] Child Obesity Facts from the National Health and Nutrition Examination Survey, *Centers for Disease Control and Prevention*, 2007-2008, www.cdc.gov.

[2] E. Fox et al. Exciting kids to walk: Enhancing walking recovery through game technology. *Games for Health Conference*, 2012.

[3] R. Kumar et al. Non-Lambertian reflectance modeling and shape recovery for faces using anti-symetric tensor splines. *IEEE Transactions on Patern Analysis and Machine Intelligence*, 2011, vol. 33(3), pp. 533-567.

[4] A. Barmpoutis et al. Beyond the Lambertian assumption: A generative model for apparent BRDF fields of faces using anti-symmetric tensor splines. *IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1-8.

[5] B. Horn, *Robot Vision*, MIT press, Cambridge, Massachusetts, 1986.

[6] O. Faugeras, *Three-Dimensional Computer Vision*, MIT press, 1993.

[7] D. Forsyth and J. Ponce, *Computer Vision: A modern approach*, 2003.

[8] A.P. Placitelli and M. Ciampi. Controller-free exploration of medical image data: Experiencing the Kinect, *24th International Symposium on Computer-Based Medical Systems*, 2011, pp. 1-6.

[9] B. Lange et al. Interactive game-based rehabilitation using the Microsoft Kinect, *IEEE Virtual Reality Workshops*, 2012, pp. 171-172.

[10] M. A. Livingston et al. Performance measurements for the Microsoft Kinect Skeleton, *IEEE Virtual Reality Workshops*, 2012, pp.119-120.

[11] B. Jian and B. Vemuri. A robust algorithm for point set registration using mixture of Gaussians, 10th IEEE International Conference on Computer Vision, 2005, vol. 2, pp. 1246-1251.

[12] V. Uriol and M. Cruz, Video-based avatar reconstruction and motion capture. California State University at Long Beach, 2005.

[13] M. C. Villa-Uriol, F. Kuester, and N. Bagherzadeh, Image-based avatar reconstruction, *In Proceedings of the NSF Workshop on Collaborative Virtual Reality and Visualization*, 2003.

[14] A. Hilton, D. Beresford, T. Gentils, R. Smith, and W. Sun, Virtual people: capturing human models to populate virtual worlds, *In Proceedings of Computer Animation*, 1999, pp. 174 –185.

[15] B. Lok, Online model reconstruction for interactive virtual environments, *In Proceedings of the 2001 symposium on Interactive 3D graphics*, ACM, 2001, pp. 69–72.

[16] S. Y. Lee, I. J. Kim, S. C. Ahn, H. Ko, M. T. Lim, and H. G. Kim, “Real time 3d avatar for interactive mixed reality,” *In Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry*, 2004, pp. 75–80.