Rapid Generation of Custom Avatars using Depth Cameras

Byoung-Keon Park † and Matthew P. Reed *†

† University of Michigan Transportation Research Institute

Abstract

Recent developments in depth-camera technology have enabled these low-cost tools to be used as body scanners. We present a simple software and hardware system using two Microsoft Kinect sensors that can generate an 3D avatar closely matching the body dimensions of an individual in about 15 seconds including scanning, post-processing and modeling time. A custom calibration of the sensors is used to improve measurement accuracy. A statistical body shape model (SBSM) is used to fit the data, thereby overcoming holes, noise, and other limitations. Pilot testing with eight child subjects shows promising results.

Keywords: Custom Avatar, Depth Camera, Statistical Body Shape Model, SBSM. Model Fitting.

1. Introduction

Recent decades have witnessed the evolution of wide range scanning technologies, including whole-body scanning systems that capture the static shape of a person to build 3D custom avatars. Most of these systems use laser-scanning technologies, but drawbacks exist for using these laser systems such as their relatively high costs and challenges with transporting and setting up the system efficiently.

Advances in depth-camera technology, which overcome some of the drawbacks of conventional laser systems, have been of particular interest in the body scanning area. Tong et al. (2012), for example, used three low-price depth cameras to scan the whole body of a subject, and then deformed a template model to reconstruct personalized avatars from the data in a few minutes. In a similar way, Weiss (2011) used SCAPE model

(a) (b)

Figure 1: Developed Kinect scanning station design: (a) Station setup for scanning with two reference poles in each side and a platform; (b) Station collapsed for transport. The reference poles, handholds, and Kinect sensors including the cables are packaged within the platform.

*Corresponding author. Email: mreed@umich.edu

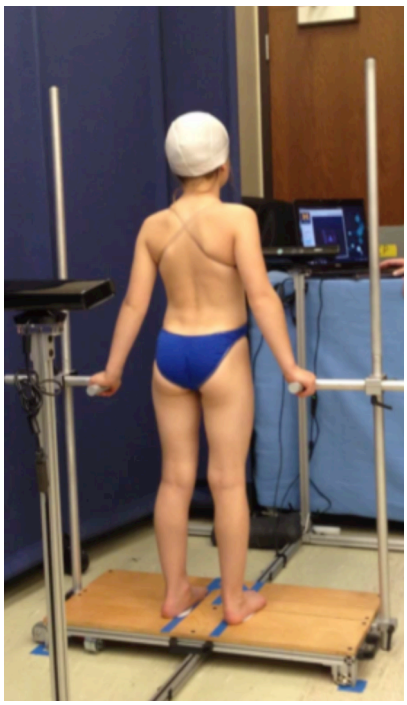


Figure 2: Subject in the scanning system.

as a template to obtain a custom avatar allowing various poses from data using depth cameras. Yu et al. (2010) presented a sub-pixel, dense stereo matching algorithm to generate realistically captured smooth and natural whole body shapes using a portable stereo vision system including 4 projectors and 8 depth cameras. However, common issues still remain in these low-cost scanning systems, such as relatively long computational time to reconstruct a complete avatar from the scan data, and the unreliability of reconstructed avatars.

This paper presents a rapid and portable scanning system using two depth cameras that can generate a reliable custom avatar. This system comprises of a portable scanning station with Microsoft Kinect sensors as depth cameras and software that builds an avatar from the scan point cloud data. The complete custom avatar is obtained by fitting a statistical body shape model (SBSM) to the scan data, thereby overcoming the common problems of data scanned using low-cost depth cameras such as holes and noise in the captured data.

2. Materials and Methods

2.1. Hardware

Figure 1 shows the system hardware, which includes two Microsoft Kinect sensors in a portable scanning station. The scanning station has a simple structure including a platform, two reference poles with handholds, and two stands for the cameras. Handholds attached to the poles help to stabilize the subjects and the poles are used for registering the multiple scan images of the subject after the scan (see below). The entire apparatus can be folded and stored in the platform, as shown in Figure 1 (b). Figure 2 shows a subject in the system, with one sensor recording data from the front and one from the back of the subject.

The Kinect utilizes an infrared emitter (IR) and an IR depth sensor, which reads the IR scatter pattern that is reflected back to the sensor. The reflected pattern is converted into depth information measuring the distance between an object and the sensor, which allows the Kinect to provide depth values of 640x480 pixels in a scene. The major design challenge associated with the scanning station primarily laid on the limitations of the Kinect specification. For example, the vertical scan range of the sensor is limited to 43 degrees, so the sensor needs to be placed relatively far from the subject to capture the whole body. However, the depth accuracy and precision of scan data are degraded as the distance increases (Andersen et al. 2012). Hence, the current system uses the integrated motor to rotate the sensor top to bottom, so that the whole body can be scanned.

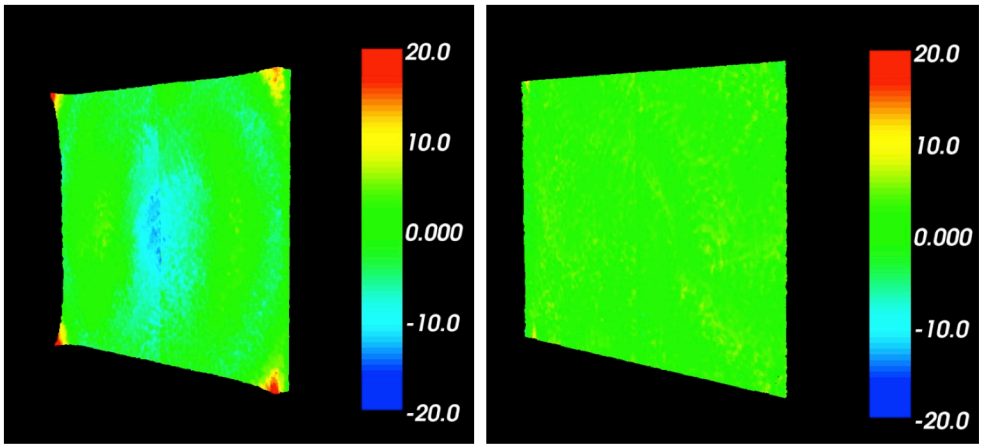
2.2. Software

2.2.1. Scanning software

A stand-alone scanning system was developed using the Microsoft Kinect SDK 1.8, which provides various useful features to obtain realistic data of the scene in real time, such as converting scan depth values to real 3D coordinates, removing background pixels and tracking joint locations. Nevertheless, acquired scan data from the sensor using the SDK produces distorted depth data and noise due to the limitations of the sensor (Andersen et al. 2012). Consequently, we developed a custom calibration to reduce these distortions.

To calibrate the camera, we first applied widely used parameters for the Kinect sensor to improve the radial and tangential distortions of the capture images (Smisek, et al. 2013). However, after scanning a flat object using the Kinect sensor, we observed that bad distortions still remain in the calibrated scan data in the normal direction of the plate, so-called depth distortion. Thus, we developed a custom calibration to improve measurement accuracy based on a 3D distortion field as follows:

1) Gather native depth values by scanning a flat object oriented perpendicular to the camera axes 10 times at 10-cm intervals (covering the scanning volume of the station).
2) Divide each scan area into 16x12 cells and average the scan depth values in each cell to reduce the noise effect, so that 16x12x10 distortion data are sampled.



(a) (b)

Figure 3: A calibration result on Kinect scan data using a custom distortion field: (a) original scan data of a flat object; (b) the calibrated result. Each pixel was colored according to the distances (mm) to a fitted plane.

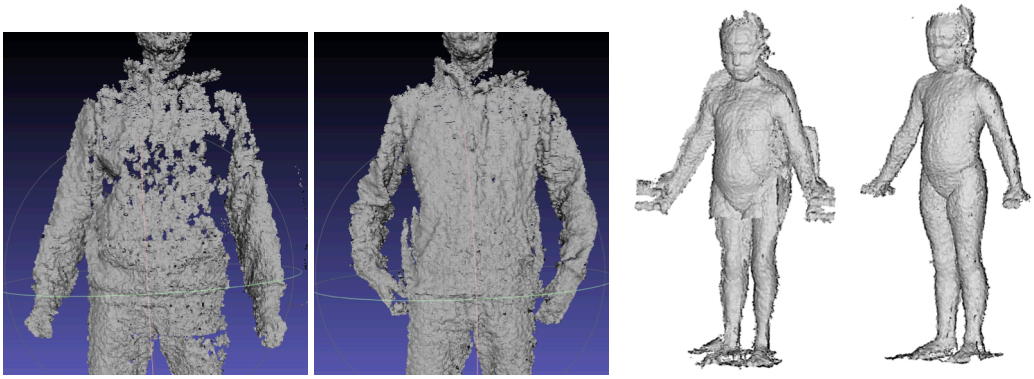
3) Build a 3D distortion field by applying a radial basis function (RBF) to the sample data.
4) Apply the distortion field to scanned depth data and then convert them to the laboratory 3D coordinates using the SDK.

Figure 3 shows a result after applying the custom calibration method to the scan data of a flat object perpendicular to the viewing direction. Each pixel in the image was colored according to the distance (mm) between the pixel point and a fitted plane. It can be observed that almost all distortions were removed from the original scan.
The next step is noise removal and hole filling. The Kinect produces significant time-varying noise in the depth data, and the sensor occasionally doesn't calculate the depth resulting in holes in the data. This problems may yield wrong fitting vectors in the next fitting step, so that the optimization performance can be degraded. To reduce the noise and holes, a total of 10 scenes were averaged from each angle (neglecting zero data). Figure 4 (a) shows a typical result from this step.

The final post-processing step is image registration. Images are captured from three angles with each of the sensors, so that a total of 6 images are obtained after averaging at each angle. The prescribed dimensions of the station such as the distance between the sensors and the angles between the scenes are used to roughly locate each image at appropriate positions and orientations. Also the geometries of the side reference poles, including the laterally oriented handles, are then used to align the images. The iterative closest point method is finally applied to each overlapped region between the images to complete the image alignment, and the final whole body data is acquired by merging the aligned data (Figure 4 (b)).

2.2.2. Rapid fitting SBSM to scan

Although Microsoft Kinect sensors have been widely used as body-scanning tools Weiss et al.



(a) (b)

Figure 4: Post-processing steps to obtain a whole body data of a subject: (a) noise and hole removal step and (b) image registration step

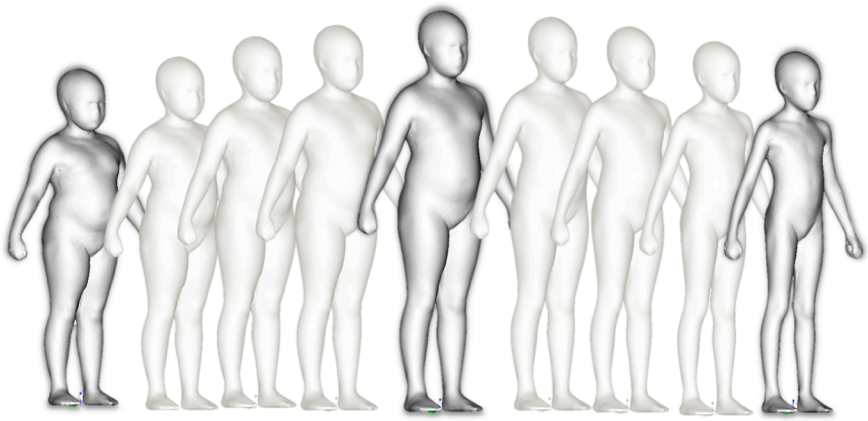


Figure 5: Morphing between individuals using the UMTRI's child SBSM. Each of the key frame models (emphasized) was generated in PCA space. The in-between models are created by linearly interpolating the PC scores of the key frames.

(2011) introduced a way to overcome many of the limitations of the data by interpreting the data using a statistical body shape model (SBSM). We have adapted a similar methodology; in essence, we search the space of body shapes for the shape most likely to have generated the observed data.

For the current project, we have employed an SBSM based on laser scan data gathered from children (Reed et al 2012, Figure 5). Standing scans of 140 children, with stature range about 100 to 160 cm and BMI 12 to 27 kg/m2, were obtained using a Vitus XXL scanner (Human Solutions). To develop a SBSM for child data, we first fit a standardized template model to each scan using a two-step process. The template model is morphed initially to match a set of 20 synthetic landmarks. An implicit surface fitting technique (Carr et al. 2001) is then applied as a fine morphing step to the roughly morphed template to represent detailed geometric features of targets. We then analyzed geometric variation of the fitted templates using principal component analysis (PCA), using an adaptation of the Allen et al. (2003) method presented by Reed and Parkinson (2008). PCA is a widely used tool to express the data on an orthogonal basis that can be more readily analyzed, and achieve data compression (Jolliffe, 2002). For the current work, we retained the first 40 PCs, which are enough to account for over 99% of the variance in the mesh coordinates

To fit the Kinect data, we first align the data to the coordinate system of the SBSM and find the value on the first PC that best matches the stature of the figure. Next, we conduct an iterative optimization to find the set of 40 PC scores that results in the best fit between the body shape generated by the SBSM and the data. The correspondence is evaluated by calculating the distance between the model mesh and the closest point in the data at 1000 points on the model mesh. Typical fitting times were less than two seconds on a typical laptop computer (Intel® Core™ i5 2.5 GHz CPU and 8 GB DDR3 RAM memory).

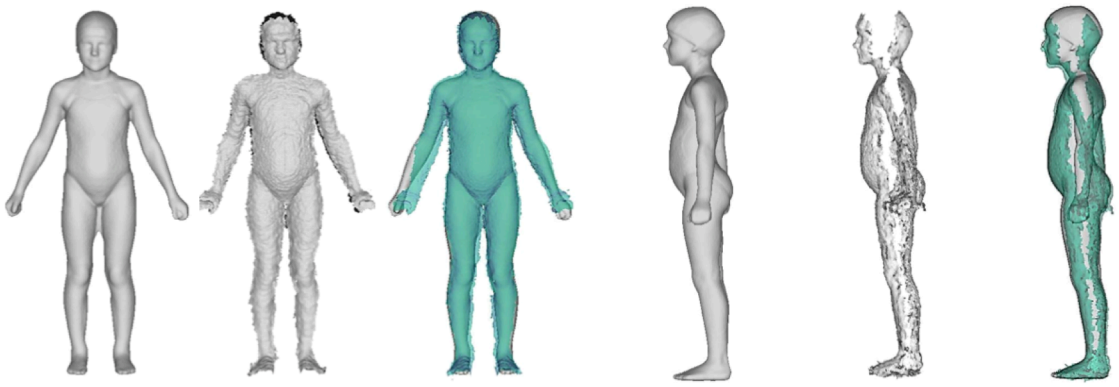


Figure 6: Sample scans of a specific subject: The left side is the scan from a Vitus laser scanner, the center image is the Kinect scan result, and the right side is the overlapped image of the two.

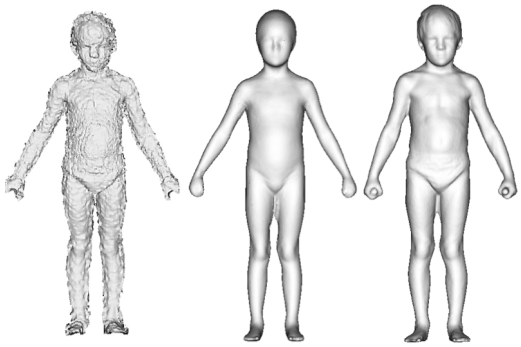


Figure 7: Example scans of a subject. A Kinect-based scan, the fitted SBSM model to the Kinect scan, and a Vitus laser scan data, respectively.

3. Results

For this pilot study, 8 children with a range of body size were scanned in both the Kinect system and the VITUS XXL laser scanner. The children wore tight-fitting swimsuits and a swim cap. During the scan, the subjects were asked to stand still in a specific posture with the arms abducted from the torso about 30 degrees and the legs slightly spread. Scanning time for each subject was 15 seconds (10 sec for scanning and 5 sec for post-processing time) Figure 6 compares a sample scan data from both the laser and Kinect systems. Although the side parts hidden from the view of the sensors were not scanned, scan data were accurate enough to measure the body shape and the anthropometric variables.

Figure 7 shows the results of fitting one the Kinect scan from one subject alongside the laser scan of the same subject. The most apparent visual difference with the laser scan is that the fitted figure is smoother and the idiosyncrasies of the face are eliminated, effectively producing an anonymized result.

To validate the feasibility of the fitted models, we quantitatively compared the models fitted to the Kinect scans, with the Vitus laser scans. The two avatars were aligned though ICP technique, and the disparity was measured in mm using the absolute distance s between the two surfaces at the model vertices. In Figure 8, the disparities are coded with the standard cold-to-hot color mapping that corresponds to 0 to 50 mm. The right side (gray image) is the Vitus scan of the same subject. The averages mean that the average of distances at all the vertices. The disparity between the two types of models was 9.44 ± 5.01 mm. Note that almost all maximum disparities were raised in extremity regions due to the differences of the subject postures between the scans.

4. Discussion

This paper demonstrated a rapid method for generating a smooth, watertight, realistic avatar from depth-camera data. The method has similarities to the Weiss et al. (2011) approach, but by imposing minimal constraints on subject posture and using two sensors we avoid many of the challenges faced in the earlier work. Our demonstration also specifically targets measurements of children.

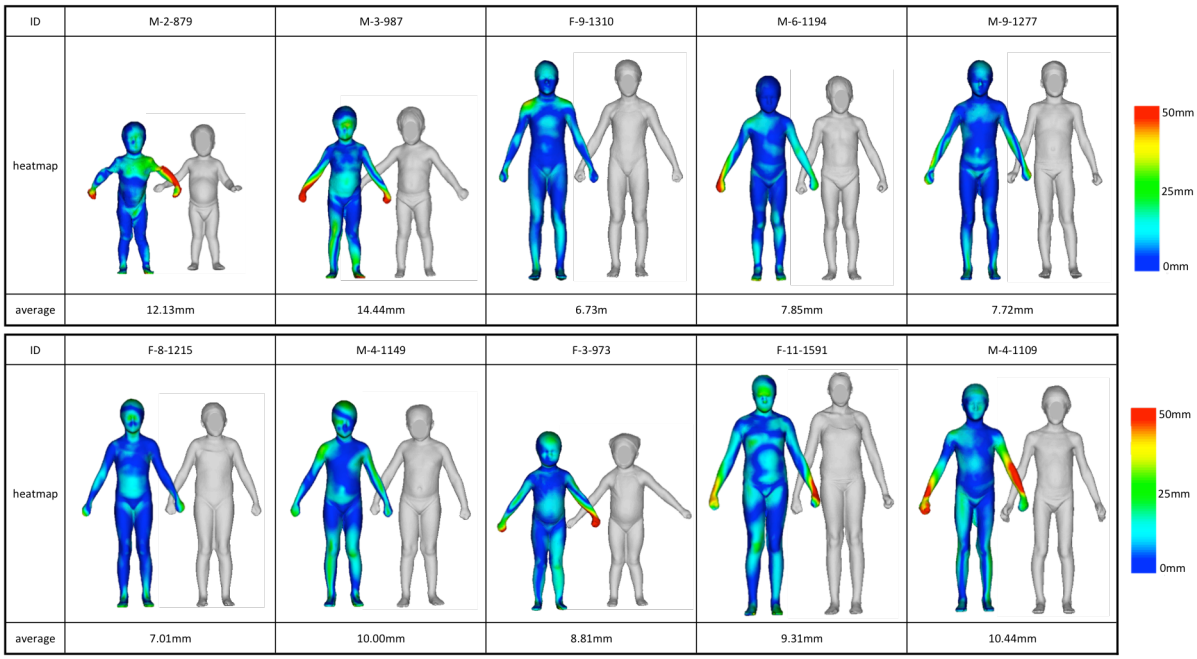


Figure 8: Sampled fitting results on child subjects. ID represents "gender-age-height in mm" of each subject. Acquired custom avatars (left side) fitted to Kinect data for subjects were colored according to the absolute distances in mm between the models and the Vitus laser scans (right side).

The pilot results demonstrate the potential for good accuracy without having complete surface data. Although various techniques have been proposed to generate a custom avatar by fitting a template model in order to overcome the drawbacks of scan data, but these approaches fill across holes and reduce noise without ensuring a realistic body shape.

In contrast, our method generating a custom avatar from a SBSM guarantees realistic body shape in areas with missing data. Besides, the template fitting approaches require many more sensors to avoid big gaps and are more sensitive to the noise in the signal. We used only two cameras and require only that our surface data not be biased, e.g., depth points systematically closer or farther from the sensor than actual. Our custom calibration and registration procedures were designed to address this. The approach is robust to noise and local errors because the SBSM does a global fit. This is analogous to spatial filtering of noise, but the filtering is based on actual human shapes rather than a simple measure of local noise.

A limitation of our method is that errors will be larger for subjects with unusual body shapes and postures, e.g., the avatar of a subject with a serious scoliosis cannot be accurately obtained since those geometric properties were not included in our SBSM model. The solution to both is to incorporate data from a wider range of postures and from people with a wider range of body shapes.

The design of the system includes several design decisions and compromises. Each Kinect is capable of capturing 30 frames per second, but our system takes approximately 10 seconds to complete a scan. This is due to the choice to place the sensors closer to the subject, thereby improving the accuracy, but necessitating images from three angles, with a delay caused by the speed of the motors that change the sensor angles. Kinect scanning systems have been demonstrated that use information gathered from the scene at high rates to localize the camera, enabling a single sensor to be moved to scan a larger area. Our system could be implemented with a single sensor, but at a cost of higher scanning time.

Acknowledgement

This research was funded in part by the MCubed program at the University of Michigan. The child statistical body shape model was developed with funding from the National Highway Traffic Safety Administration. The hardware system was developed by a team of engineering students at the University of Michigan: Ryan Holstad, Bret Kirchner, John Lavoie-Mayer, and Nathan Van Nortwick.

References

Andersen, M. R., Jensen, T., Lisouski, P., Mortensen, A. K., Hansen, M. K., Gregersen, T., & Ahrendt, P. (2012). Kinect depth sensor evaluation for computer vision applications. Århus Universitet.

Carr, J.C., Beatson, R.K., Cherrie, J.B., Mitchell, T.J., Fright, W.R., McCallum, B.C., and Evans, T.R. (2001). Reconstruction and representation of 3D objects with radial basis functions. SIGGRAPH 01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques.

Jolliffe, I. (2002). Principal Component Analysis, 2 ed. Springer.

Reed, M.P. and Parkinson, M.B. (2008). Modeling Variability in Torso Shape for Chair and Seat Design. Proceedings of the ASME Design Engineering Technical Conferences.

Reed, M.P. (2012). A pilot study of three-dimensional child anthropometry for vehicle safety analysis. Proceedings of the 2012 Human Factors and Ergonomics Society Annual Meeting. HFES, Santa Monica, CA.

Smisek, J., Jancosek, M., & Pajdla, T. (2013). 3D with Kinect. In Consumer Depth Cameras for Computer Vision (pp. 3-25). Springer London.

Tong, J., Zhou, J., Liu, L., Pan, Z., & Yan, H. (2012). Scanning 3d full human bodies using kinects. Visualization and Computer Graphics, IEEE Transactions on, 18(4), 643-650.

Yu, W., & Xu, B. (2010). A portable stereo vision system for whole body surface imaging. Image and vision computing, 28(4), 605-613.

Weiss, A., Hirshberg, D., & Black, M. J. (2011). Home 3D body scans from noisy image and range data. In Computer Vision (ICCV), 2011 IEEE International Conference on (pp. 1951-1958). IEEE.