Accurate 3D Face and Body Modeling from a Single Fixed Kinect

Ruizhe Wang*, Matthias Hernandez*, Jongmoo Choi, Gérard Medioni
Computer Vision Lab, IRIS
University of Southern California

Abstract

In this paper, we address the problem of both face and body modeling using a single fixed low-cost 3D camera (e.g. Kinect). Unlike other scanning technologies which either set up multiple sensors around a static subject or scan the static subject with a hand-held moving 3D camera, our method allows the subject to move in front of a single fixed 3D sensor. This opens the door to a wide range of applications where scanning is conveniently performed at home alone. While partial range scans of face are aligned rigidly, the body modeling is performed through articulated registration. At the surface reconstruction stage, we utilize the cylindrical representation for smoothing, hole filling and blending. Experimental results demonstrate the effectiveness of our modeling technique.

Keywords: Kinect, face modeling, body modeling, rigid registration, articulated registration, cylindrical representation

1. Introduction

3D face and body modeling is of interest to computer vision and computer graphics. Precise 3D face and body models are necessary in many applications, such as animation, virtual reality, human computer interaction. However, obtaining such an accurate model is not an easy task. Early systems are either based on laser scan or structured light. While these systems can provide very accurate models, they are expensive. In this paper, we propose a user-friendly scanning system based on a low-cost structured -light 3D sensor.

The complete setup of our home-used scanning system is illustrated in Fig. 1. The camera is mounted vertically to maximize the field of view so that a subject can stand as close as possible. The subject first scan the body. The required initial pose of the subject is also shown in Fig. 1. While the subject starts turning from the initial pose, he/she is required to stay static at approximately every 1/4 circle for 1 second. The subject can turn naturally as long as his/her two arms stay in the torso's plane and his/her two arms do not cause occlusion on legs. After scanning the body, the user comes closer to the sensor and scans his/her face. While the system and instructions are easy to set up and follow respectively, the whole data-recording process won't take more than 20 seconds.

Fig. . System setup

The first key component of our scanning technology is the registration between point clouds. For face scanning, we perform rigid registration. We set a frontal depth image as the reference, and then align each subsequent cloud of 3D points to the reference using a GPU (Graphics Processing Unit)

* equal contribution to the paper.
contact: ruizhewa@usc.edu; (626)390-8301

implementation of a robust variant of the ICP (Iterative Closest Point) algorithm, which enables real time scanning. This registration, combined with our online surface reconstruction method, allows us to reject poor alignment due to facial expressions, occlusions, or a poor estimation of the transformation, by thresholding the distance between the current frame and our online model. For body modeling, we carry out articulated registration considering the subject turns 360 degrees in front of the sensor during scanning. Instead of registering all consecutive frames, we sample key frames out of the complete turning sequence and align them in a top-bottom-top manner.

The second major component is surface reconstruction. For face modeling, we utilize an online cylindrical representation which enables us to perform 3D operations in the unwrapped 2D image space. For example, interpolation on a 2D image is hole filling on 3D surface and filtering the image acts as spatial smoothing in 3D. We initialize the cylindrical model with the reference point cloud and update the model with reliable new frames. Following the same idea, we employ a part-based cylindrical representation for the body modeling and solve the problem of blending between two overlapping cylinders. After separately scanning face and body, we register and blend them to generate a single model.

Experimental results demonstrate the effectiveness of our modeling technique. For both face and body modeling, we quantitatively compare our models to commercial laser scans. For the face, an average error of 1mm is achieved while for the body an average error of 5mm is observed. The degradation of accuracy for the body modeling is mainly due to the sensor's increasing quantization step at a longer distance. With more accurate 3D sensing device, better accuracy can be achieved.

The rest of the paper is organized as follows. Section 2 covers our face and body modeling methods in details. Section 3 describes the experimental results. Section 4 ends with conclusion and future work.

2. Method

2.1. Face modeling

Reconstructing an accurate 3-D model from the 3-D camera is not an easy problem. The depth map computation uses a triangulation principle, hence making the depth data near boundaries very noisy. As a result, a simple averaging in time does not suffice.

We propose to accumulate and register face images with several poses. Then, we perform both a temporal integration and a spatial smoothing to remove the noise. A cylindrical representation [3,4] is used to improve performance. Fig. 2 shows a flowchart of our approach. This approach was presented in [5].

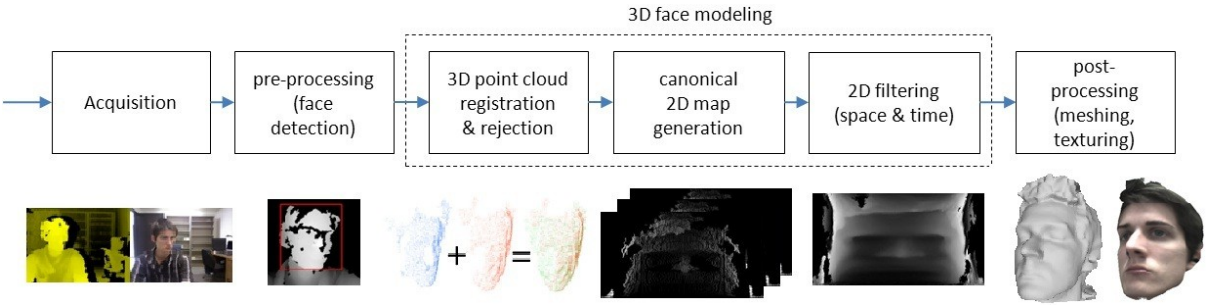

Fig. . Flowchart for the face modeling

After segmenting the face region, the images are aligned by using a point cloud registration technique, namely the Expectation-Minimization Iterative Closest Point algorithm [1,2]. The first image is arbitrarily set as a reference and assumed to be a frontal face. The subsequent images are

registered to the reference coordinate system. Note that other pose estimation algorithms could be used instead, as long as they provide sufficient accuracy.

The registered images are merged in an unwrapped cylindrical 2-D image, which can represent star-shaped objects, such as faces [3,4]. Practically, a cylinder is set around the face in the reference frame. Its axis is the vertical axis going through the middle of the head. The axis location does not need to be very accurate and is loosely estimated by taking the center of mass of the head in the reference frame. For each 3-D point of coordinates (x, y, z), the cylindrical coordinate (✕, ¯, y) can be computed with the equations:

$$\rho = \sqrt{x^2 + z^2},$$

$$\theta = \begin{cases} 0 & \text{if } (x = 0 \wedge y = 0) \\ \arcsin \dfrac{z}{\rho} & \text{if } x \geq 0 \\ \pi - \arcsin \dfrac{z}{\rho} & \text{if } x < 0. \end{cases}$$

The 3-D geometry of the face can be stored in an unwrapped cylindrical image in which the value at the pixel (¯, y) is the distance ✕ from this point to the cylinder axis (Fig. 3). This representation is simple and enables us to perform 2-D image filtering techniques to smooth the 3-D data, hence improving performance. Also, the mesh can be computed very simply by linking neighboring pixels together.

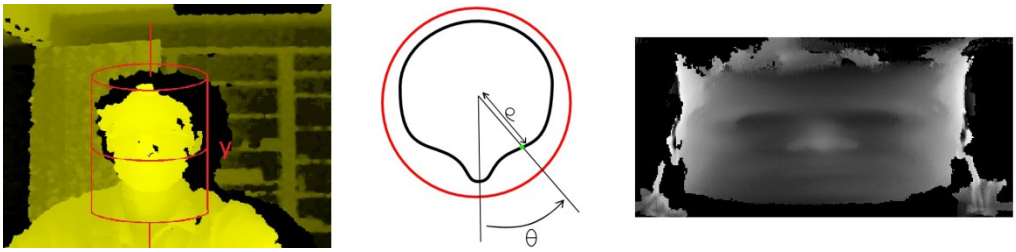

Fig. . Cylindrical representation. A cylinder is set around the face (left). Every 3-D point (x; y; z) is converted to cylindrical coordinate (✕, ¯, y) where r is the distance from the surface to the cylinder axis (middle), giving the intensity value at the pixel (¯, y) in the unwrapped cylindrical map (right).

2.2. Body modeling

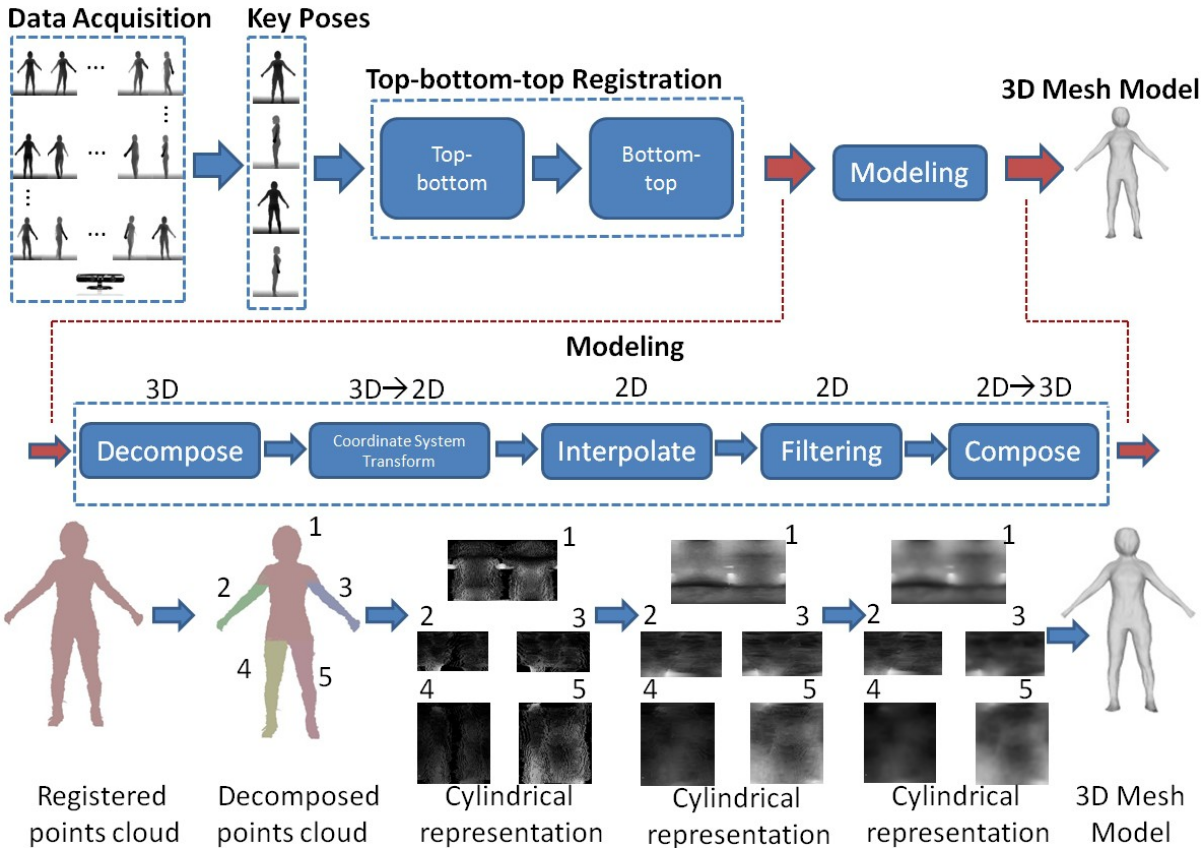

Fig. . Flowchart for the body modeling

The general pipeline of the body modeling system is shown in Figure 4. First we detect 4 key poses out of the whole depth video sequence, which are front (reference pose), back, and two profiles. They cover sufficient information to describe the main body shape. The 4 key frames are registered in a top-bottom-top manner instead of other well-developed non-rigid registration methods. Top-bottom means that registration goes from the root node to all leaf nodes in the tree model of human body as shown in Figure 1(c) while bottom-top means the opposite. Top-bottom registration first aligns the torso or the whole body then aligns succeeding rigid body parts. Bottom-top registration first refines the alignment of rigid body parts and then propagates the refinement all the way to the root node, i.e. torso. After registering 4 key frames, an articulated part-based cylindrical body model, which supports a set of operations, can be used to process the rough and noisy registered points cloud of the body. Figure 4 shows a flow chart of the modeling process. The key here is the 2D part-based unwrapped cylindrical representation which enables computationally effective 2D interpolation and 2D filtering. More details can be found in [6].

2.3. Blending face and body models

The performance of the Kinect sensor decreases quadratically as the distance increases. At the scanning range of full body, e.g. approximately 2 meters, most details are lost while we only capture the global shape. Among all those lost details, the ones of face are most appealing. Hence it is necessary to blend the body model and face model to take advantage of the close scanning distance of face.

After rigid and articulated registration stage, we obtain the raw points cloud of face and body respectively. We rigidly align the points cloud of face with the points cloud belonging to the head area of body. Then we remove all points, which are closer than a threshold to the facial points cloud, from

the body. This makes sure that at the surface reconstruction stage, only the points cloud of face is used for the final model's facial area. The gap between the face points cloud and body points cloud are filled automatically by our cylindrical surface reconstruction technique.

3. Experimental Result

3.1. Face models and evaluation

The 3-D face models can be reconstructed online and in real-time. Also, they are visually accurate (Fig. 5). The accuracy was quantified by comparing our models to commercial laser scans and we could get an accuracy of 1mm on average. Note that smooth areas such as the cheeks are very close to the ground truth while the high curvature areas have a bigger error, as shown in Fig. 6.

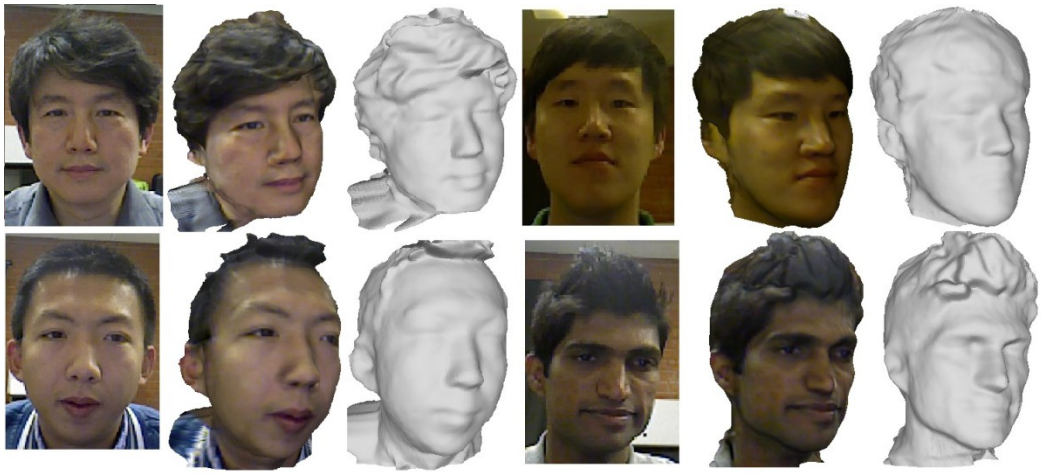

Fig. . Face models of different people

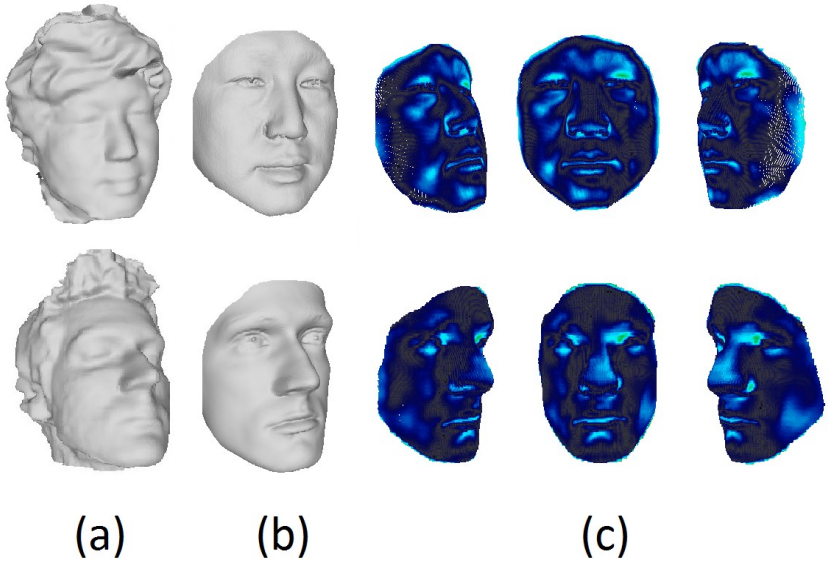

Fig. . (a) Our models (b) Laser scans (c) Heat map

3.2. Body models and evaluation

Fig. 7 includes modeled results of 4 people scanned by our system. When they turn in front of the depth camera, we observe obvious articulated motion. In Fig. 7 each model is showed at 4 different views. Fig. 7 shows clear and smoothed body shapes as a whole and contains personalized shapes such as knees, hips and clothes. The holes on body are interpolated and the joints between rigid body parts are well blended. The average computing time of the whole process with an Intel Core i7 processor at 2.0 GHz is around 3 minutes. Although finer details of body shape (e.g. lips, eyes)

cannot be extracted due to the noisy nature of Kinect, we believe that the current system can generate models accurate enough for applications such as online shopping and gaming. We believe that the model can be further refined by using a more complex model, adding more frames and taking advantage of the corresponding RGB image.

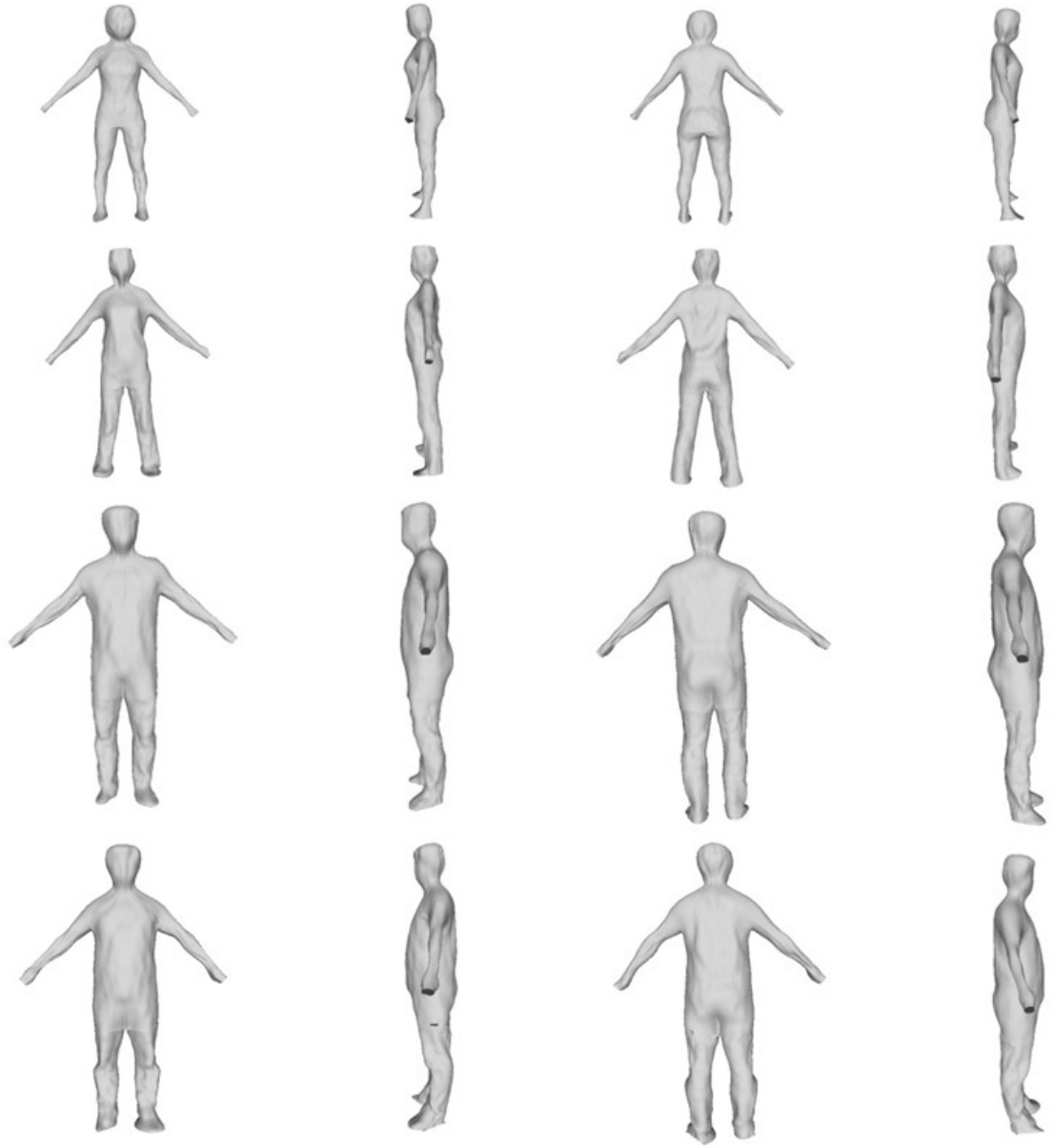

Fig. . Body models of different people

Besides qualitative analysis, we also present quantitative comparison between our model and the laser scanned result. Due to the existence of articulation between these two models, it is hard to compare them as a whole. Instead we compare the segmented rigid body parts. The heat map of torso is shown in Fig. 8. We generate the point-wise error by mapping each point of our body part to the cylindrical image generated by the laser-scanned body part and looking for the closest pixel or interpolating neighboring 4 pixels. The median of absolute error on torso is 5.84mm.

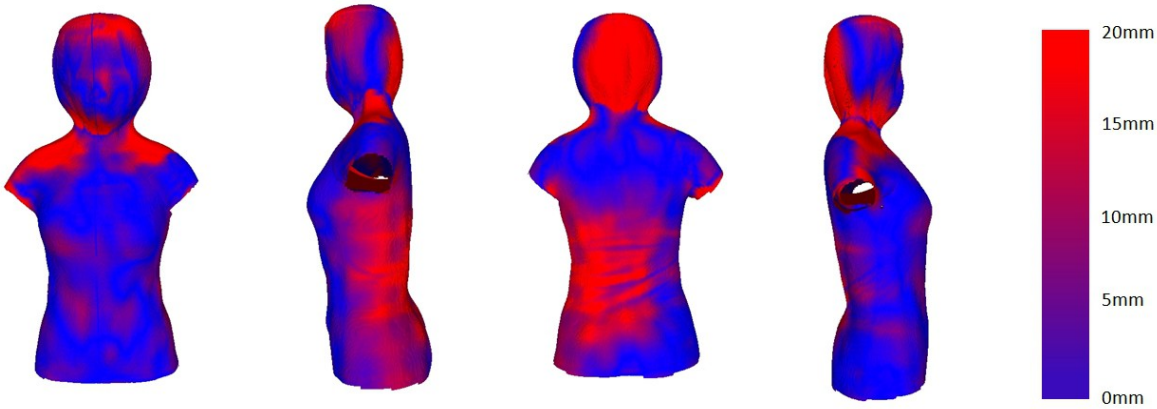

Fig. . Heat map of torso

3.3. Blending face and body models

We manually segment the head points cloud from body to register with the face points cloud (Sec 2.3). The final reconstruction is shown in Fig. 9. We capture the global shape of body as well as the fine details of face.

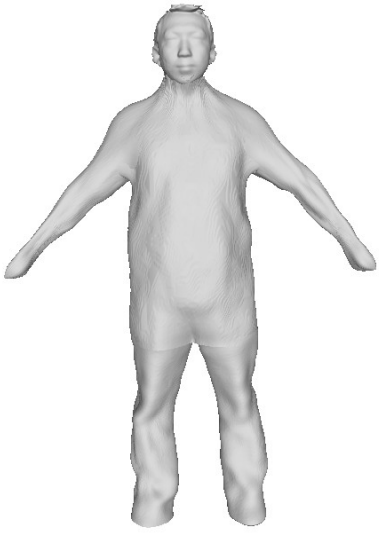

Fig. . Final reconstruction model

4. Conclusion and Future Work

In this paper, we address the problem of both face and body modeling with a single fixed 3D sensor. Our scanning system generates high-quality face and body models quickly. In the future, we plan to work on refining details of the model from the intensity stream.

References

[1] Chen, Y; Medioni G: "Object modelling by registration of multiple range images". *Image Vision Computing, 1991*

[2] Tamaki T, Abe M, Raytchev B, Kaneda K: “Softassign and EM-ICP on GPU”. CVPR 2010.

[3] Lin Y, Medioni G, Choi J: “Accurate 3D face reconstruction from weakly calibrated wide baseline images with profile contours”. CVPR 2010.

[4] Williams L: “Performance-Driven Facial Animation”. Computer Graphics 1990, 24(4).

[5] Hernandez M, Choi J, Medioni G: “Laser scan quality 3-D face modeling using a low-cost depth camera”. European Signal Processing Conference (EUSIPCO) 2012.

[6] Wang R, Choi J, Medioni G.: "Accurate Full Body Scanning from a Single Fixed 3D Camera" 3DIMPVT, 2012 Second International Conference on. IEEE, 2012: 432-439.