Accurate Full Body Scanning from a Single Fixed 3D Camera

Ruizhe Wang Jongmoo Choi Gerard Medioni
Computer Science Department, University of Southern California
3737 Watt Way, PHE 101, Los Angeles, CA, 90089
{ruizhewa, jongmooc, medioni}@usc.edu

Abstract

3D body modeling has been a long studied topic in computer vision and computer graphics. While several solutions have been proposed using either multiple sensors or a moving sensor, we propose here an approach when the user turns, in a natural motion, in front of a fixed 3D low cost camera. This opens the door to a wide range of applications where scanning is performed at home. Our scanning system can be easily set up and the instructions are straightforward to follow. We propose an articulated, part-based cylindrical representation for the body model, and show that accurate 3D shape can be automatically estimated from 4 key views detected from a depth video sequence. The registration between 4 key views is performed in a top-bottom-top manner which fully considers the kinematic constraints. We validate our approach on a large number of users, and compare accuracy to that of a reference laser scan. We show that even using a simplified model (5 cylinders) an average error of 5mm can be consistently achieved.

1. Introduction

3D body modeling is of interest to computer vision and computer graphics. A 3D precise body model is 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.

Image-based approaches play an important role in body modeling. Shape-from-silhouettes (SFS) method [13, 14] can give us very good result given many synchronized cameras. The advantage is that only the silhouette information is used hence we do not need to pay extra attention to take good care of textures. The disadvantage, on the other hand, is also obvious: Accuracy heavily relies on the number of synchronized cameras which restrict its application in real life.

The advent of a new type of range sensors, e.g. Microsoft Kinect [12], has drawn significant attention in computer graphics and computer vision. Several methods and two commercial systems (*i.e.* Styku [22] and Bodymetrics [2]) based on the depth camera have been proposed for body modeling. These methods can be generalized and categorized in the following scenarios: *1) multiple fixed sensors with static person; 2) multiple fixed sensors with moving person; 3) single moving sensor with static person; 4) single fixed sensor with moving person.* Our main focus in the paper is to establish a convenient home-used body modeling system so that a naive user can easily scan his/her body alone. This requirement leaves the 4th option as the only viable one. So we try to address 4) in this paper, i.e. single

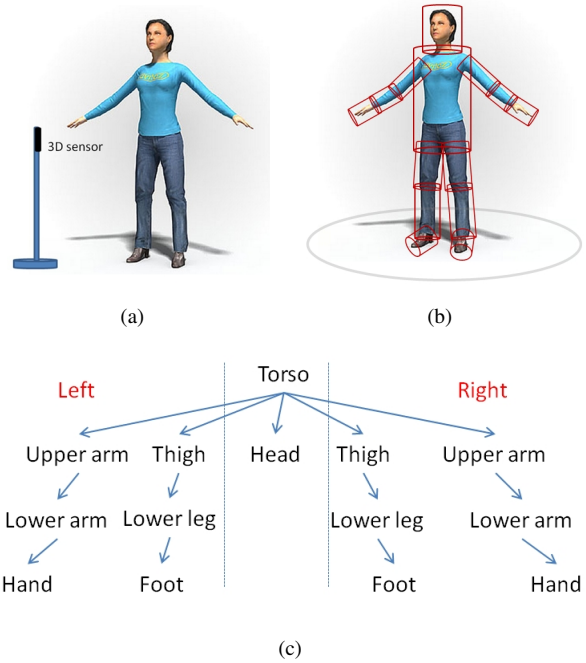


Figure 1: (a) System setup (b) Articulated part-based cylindrical representation of human body (c) Tree representation of human body

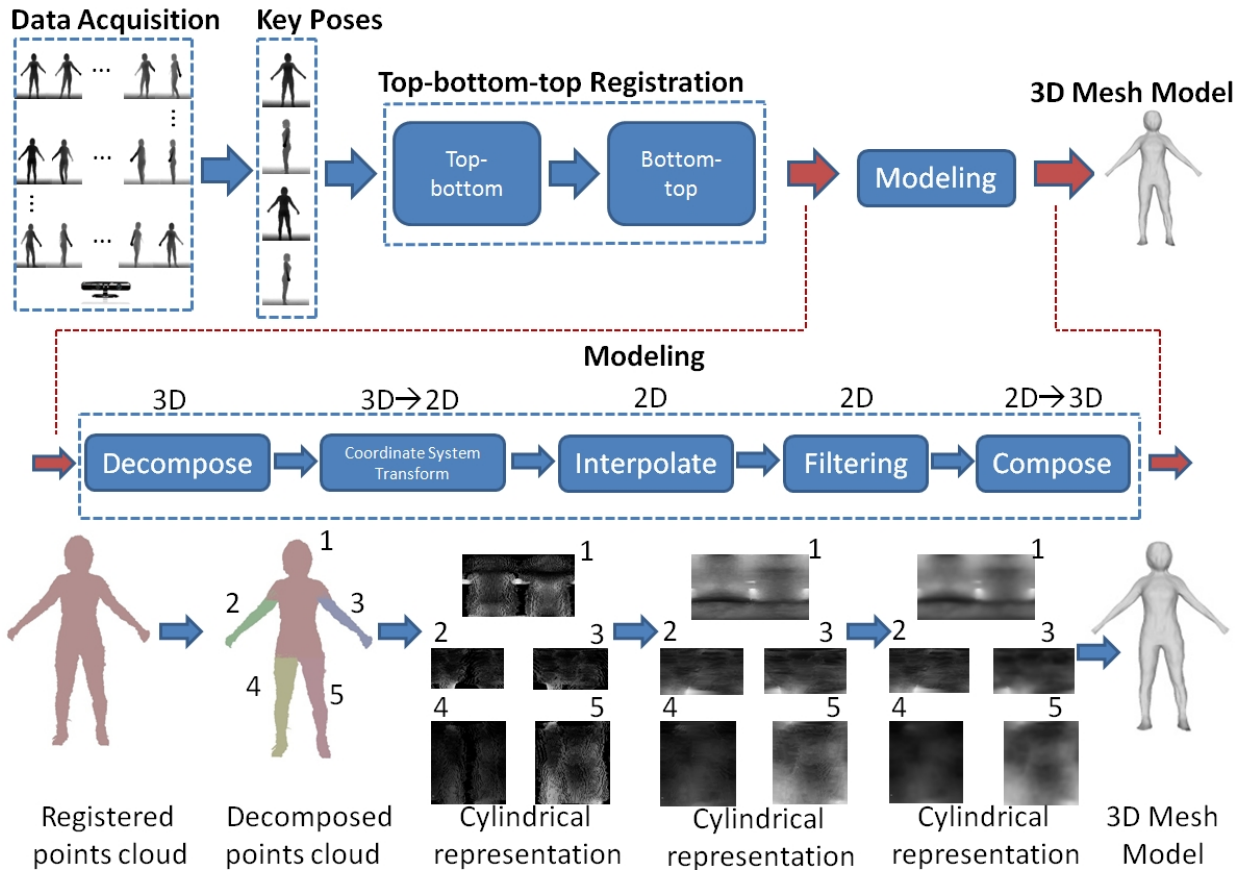


Figure 2: General pipeline of our scanning system

fixed sensor with moving person.

The complete setup of our home-used scanned system is illustrated in Figure 1(a). The camera is mounted vertically to maximize the field of view so that a subject can stand as close as possible. The required initial pose of the subject is also shown in Figure 1(a). 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. While the system and instructions are easy to set up and follow respectively, the whole data-recording process won't take more than 10 seconds.

This user-friendly scenario imposes serious algorithmic challenges. **Articulation:** Articulation needs to be considered even when the person is turning on an automatic turntable [4, 26]. Not to mention when the person is turning in a nature way in front of the Kinect sensor. The more cylinders we use in the model, the more degree of freedom (DOF) we need to estimate. **Noisy nature of Kinect:** While Kinect can obtain reasonable results at close distances, its behavior decreases quadratically as we go further away which is verified in [11]. Our normal working distance is 2m, and this gives us an statistical error $\varepsilon \sim N(0, \sigma)$ where σ is approximately 7mm. **Lack of texture and accurate silhouette information:** With only range images, we lose access to the texture information. Big variance on boundary of object in range images makes silhouette noisy. So in our paper we cannot use the two most powerful hints to register and refine the model.

In this paper, we propose a body modeling system with a single fixed range camera which can be used conveniently at home. The general pipeline of the working system is shown in Figure 2. We first detect 4 key poses out of the whole depth video sequence, which are front (reference pose), back, and two profiles. The 4 key frames are registered in a top-bottom-top manner. 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 propagate refinement all the way to the root node, *i.e.* torso. After registering 4 key frames,

an articulated part-based cylindrical body model (as shown in Figure 1(b)), which supports a set of operations, can be used to process the rough and noisy registered points cloud of the body. Figure 2 shows a flow chart of the modeling process. The key here is the 2D part-based cylindrical representation which enables computationally effective 2D interpolation and 2D filterings.

The contributions of this paper can be summarized as follows: 1) A body scanning system which can be conveniently used at home by a single naive user; 2) A new articulated part-based representation of 3D body model which supports a set of operations, such as composition, decomposition, filtering, and interpolation; 3) A simple method to automatically detect 4 key poses from a sequence of range images; 4) A top-bottom-top registration between 4 key frames; 5) A quantitative evaluation of depth quality.

The rest of the paper is organized as follows. Section 2 covers related state-of-the-art algorithms. Section 3 details our proposed body model, i.e. the articulated, part-based cylindrical representation and a set of operations that it supports. In section 4, we give our top-bottom-top registration process. Section 5 includes our experimental results and a quantitative comparison analysis with a laser scanned result. We conclude briefly with Section 6.

2. Related Work

At a first look, this is a non-rigid registration problem on which a huge amount of work has been proposed. [6] extends the iterative closest point algorithm (ICP) [30] by softening hard links between points and formulating the registration process under a probabilistic scheme. Embedded deformation model [23] proved to be successful in non-rigid registration [8, 15, 26]. A similar approach [16] retains smoothness of warped surface by using Laplace-Beltrami operator. All those methods require either enough overlapping areas between consecutive frames or accurate range scans as inputs. In our case, however, we are registering four noisy frames with barely overlapping area which restricts the application of the well-developed non-rigid registration methods in our scenario.

Several related works exist on body modeling. They can be roughly classified as four main categories based on the number of cameras used and whether the camera or the person is moving. Different scenarios make different underlying assumptions and hence lead to different research focus.

Multiple Fixed Cameras with Static Person. Before the introduction of the nowadays popular range sensors (e.g. Kinect), people were able to obtain full model from several fixed intensity cameras by Shape-From-Silhouette (SFS) algorithm [14, 19]. The modern SFS-based approaches use surface-based representation which allows to use regularization in an energy minimization framework and give pretty impressive results [13]. Currently, after more and more range cameras hit the market, researchers have been working on 3D modeling with this type of sensor since it can directly give you the shape of the object. Commercial systems are using multiple calibrated depth cameras to model body shape. The person is required to stay static during procession and it is quite straightforward to align several points clouds from different cameras to get the final model. Both methods can provide the most accurate model so far. However, they both require multiple synchronized cameras and the later even needs to deal with interference between cameras [18]. These factors make them far from home applications.

Multiple Fixed Cameras with Moving Person. Motion of a person can greatly decrease the number of required fixed cameras. Because provided a good registration result between consecutive frames, it is exactly the same as setting up more cameras. For the SFS-based approaches, the registration is achieved by locating the Colored Surface Points (CSP) [3, 4]. The corresponding CSP between two consecutive frames incorporates the 6 DOF rigid motion information. For the Depth-based approaches, a good registration can be provided either by articulated version of the Iterative Closest Point (ICP) algorithm [30] or by iteratively performing pair-wise non-rigid geometric registration and global registration [26]. Again, these methods still require 3 or 4 synchronized cameras and the person must be standing on a turntable and try to remain rigid.

Single Moving Camera with Static Person. SFS-based approach only works when the camera is mounted on a robot arm and moves circularly around the static body. While under circular motion, the DOF of fundamental matrix between consecutive frames is restricted to be 4 [28]. Parameters of all fundamental matrix then can be estimated by minimizing the reprojection errors of corresponding epipolar tangents [29]. And the model of the body can be reconstructed from the motion estimated from fundamental matrixes. Although the result is quite impressive, the system set up is still impractical. As for Depth-based approach, KinectFusion [9] proves to be a great success. Based on dense tracking and volumetric representation, it can reconstruct a static 3D scene in real time. While it works on a static scene, it fails when registering only the points from a moving person in the presence of articulation.

Single Fixed Camera with Moving Person. The only proposed approach [27] so far is depth-based, and uses statistical learning. [27] estimated the body shape by fitting image silhouettes and depth data to the Shape Completion and Animation of People (SCAPE) model [1]. The work can be understood under a learning scheme where the best model is predicted by the observed data. The final model is a best match among all candidates of a limited subspace instead of being generated from the data itself. The bias introduced by the learning scheme is inevitable. Moreover, the

whole system takes approximately 65 minutes to optimize, which is too slow for practical applications. Our focus is also on single fixed camera with moving person, we align the detected 4 key frames completely based on the prior geometry information.

3. Generic Human Body Model and Operations

The generic body model used in this paper and its supported operations are important for both top-bottom-top registration and modeling.

3.1. Generic Human Body Model

The generic human body model is depicted in Figure 1(b). The body model consists of a set of cylinders representing rigid body parts and respects kinematic constraints,*i.e*. each rigid part is connected with its parent via a joint. This model also has a tree representation (Figure 1(c)) which is important for understanding our top-bottom-top registration. In the tree representation, each node is a rigid body part and each edge is a joint. Assuming that we have n nodes and n edges, each node B_i and each edge J_i can be represented as

$$B_i = \{\mathbf{P_i}, \mathbf{I_i^{J_i}}\}, i = 1...n, \tag{1}$$

$$J_i = \{\vec{j}_{ic}, (\hat{j}_{ix}, \hat{j}_{iy}, \hat{j}_{imain})\}, i = 1...n. \tag{2}$$

where J_i and B_i forms a pair and J_i connects B_i to its parent node. Each joint J_i has four vectors which constitute the local Cartesian Coordinate System. $\vec{j}_{ic}$ is the location of joint i and the origin of local system. $\hat{j}_{imain}$ is the main direction of cylinder i and the z axis of local system. $\hat{j}_{ix}$ and $\hat{j}_{iy}$ are the x axis and y axis of the local system. A specific body part B_i has two components. $\mathbf{P}$ is the points cloud, which can be represented in the World Cartesian Coordinate System as $\mathbf{P_i^W}$, in the Local Cartesian Coordinate System defined by J_i as $\mathbf{P_i^{L_i}}$ or in the Local Cylindrical Coordinate System defined by J_i as $\mathbf{P_i^{C_i}}$. $I_i^{J_i}$ is the cylindrical image and it can also be regarded as a compact and discretized version of $\mathbf{P_i^{C_i}}$.

3.2. Supported Operations

Decomposition $(\mathbf{3D})$. Given a points cloud $\mathbf{P_{body}^W}$ of a whole human body, we decompose it and obtain $\mathbf{P_i^W}$ of each B_i in (1) by taking advantage of planar geometry. Assuming that accurate critical points information is given (section 4.2) we can easily decompose the whole body into several rigid parts. A 2D example decomposing crotch is given in figure 3. The whole body can be decomposed in a similar way. Notice that after decomposing, connected rigid body pair has an overlapping area, *i.e*. $\mathbf{P_i^W} \bigcap \mathbf{P_j^W} \neq \phi$ if B_i and B_j are connected by joint J_i. The overlapping area is useful for blending between connected body parts when we compose everything into a whole model.

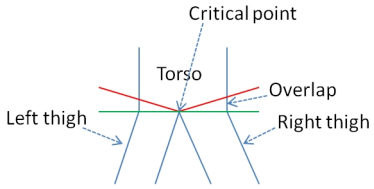


Figure 3: Decomposition at crotch

Local Structure Extraction $(\mathbf{3D})$. After decomposition, every point in $\mathbf{P_{body}^W}$ is assigned to a rigid body part B_i. We use Pincipal Component Analysis (PCA) [10] to extract the structure $(\hat{j}_{ix}, \hat{j}_{iy}, \hat{j}_{imain})$ of each rigid part. The largest component defines the axis of the corresponding cylinder and hence is $\hat{j}_{imain}$. The order of $\hat{j}_{ix}$ and $\hat{j}_{ix}$ does not matter as long as $(\hat{j}_{ix}, \hat{j}_{iy}, \hat{j}_{imain})$ forms an Cartesian Coordinate System in space. Notice that during the decomposition process, the axis information is used. So in practice we apply decomposition and local structure extraction iteratively until all the axes converge.

Global and Local Cartesian Coordinate System Transformation $\mathbf{P_i^W} \rightarrow \mathbf{P_i^{L_i}}(\mathbf{3D} \rightarrow \mathbf{3D})$. For each rigid body part B_i, we need to represent its points cloud $\mathbf{P_i^W}$ in the Local Cartesian Coordinate System obtained from decomposition and local structure extraction. Given the basis $(\hat{j}_x, \hat{j}_y, \hat{j}_{main})$ which are the new basis and the center $\overrightarrow{j_c}$. Then for any point $\overrightarrow{p_{ij}}^w \in \mathbf{P_i^W}$ we have the transformation $\overrightarrow{p_{ij}}^{l_i} = \mathbf{R_i}\overrightarrow{p_{ij}}^w + \overrightarrow{t_i}$ where $\mathbf{R_i} = [\hat{j}_{ix}, \hat{j}_{iy}, \hat{j}_{imain}]^{-1}$ and $\overrightarrow{t_i} = -\mathbf{R_i}\overrightarrow{j_{ic}}$.

Local Cylindrical Image $\mathbf{P_i^{L_i}} \rightarrow \mathbf{P_i^{C_i}}(\mathbf{3D} \rightarrow \mathbf{2D})$. After mapping each point to its corresponding Local Cartesian Coordinate System $\overrightarrow{p_{ij}}^{l_i} = [p_{ijx}^{l_i}, p_{ijy}^{l_i}, p_{ijz}^{l_i}]$, it can be further transformed into the Cylindrical Coordinate System $\overrightarrow{p_{ij}}^{c_i} = [\rho_{ij}^{c_i}, \varphi_{ij}^{c_i}, z_{ij}^{c_i}]$ where $\rho_{ij}^{c_i} = \sqrt{{p_{ijx}^{l_i}}^2 + {p_{ijy}^{l_i}}^2}$, $\varphi_{ij}^{c_i} = \arccos(\frac{p_{ijx}^{l_i}}{\rho_{ij}^{c_i}})$ and $z_{ij}^{c_i} = p_{ijz}^{l_i}$. $\overrightarrow{p_{ij}}^{c_i}$ can be further discretized and mapped to $\mathbf{I_i^{J_i}}$ [7].

Interpolation $(\mathbf{2D})$. Interpolation is a main reason for our cylindrical representation. In many cases even after we gather many frames and use all the points, we still have holes on the cylindrical image due to occlusions. We propose linear interpolation along rows of the cylindrical image which turns out to be circular interpolation in space.

Filtering $(\mathbf{2D})$. Filtering is another reason for us to use the cylindrical representation. After mapping 3D points to a 2D cylindrical image, we can easily take advantage of the well-developed 2D image filters. We use the bilateral filter [25] for spatial smoothing which can reduce noise while

preserving edges. For temporal filtering, we use the running mean on each pixel on the cylindrical image [7].

Composition ($\mathbf{2D} \rightarrow \mathbf{3D}$). We compose single cylinders to construct the 3D mesh body model. The gap between connected rigid body parts must be blended. The 2D cylindrical images of connected body parts overlaps (this is not reflected in Figure 2). Assuming B_i is the parent of B_j, we can decompose the overlapping points cloud $\mathbf{P_{ij}^W}$ out of $\mathbf{P_i^W}$ using the same planar geometry defined before. $\mathbf{P_{ij}^W}$ is mapped in the following directions: $\mathbf{P_{ij}^W} \rightarrow \mathbf{P_{ij}^{L_j}} \rightarrow \mathbf{P_{ij}^{C_j}} \rightarrow \mathbf{I_{ij}^{J_j}}$. We linearly blend $\mathbf{I_j^{J_j}}$ with $\mathbf{I_{ij}^{J_j}}$ to update $\mathbf{I_j^{J_j}}$. Connectivity on 2D cylindrical image help us mesh rigid body part easily.

Algorithm 1 Articulated registration based on EM-ICP

Input: A reference decomposed points cloud $\mathbf{Q^{ref}} = \{\mathbf{P_i^{ref}}, i = 1...n\}$, a points cloud $\mathbf{Q^{in}}$.
Output: A decomposed and registered points cloud $\mathbf{Q^{out}} = \{\mathbf{P_i^{out}}, i = 1...n\}$.

```
Register Q^in with Q^ref as a whole by EM-ICP
Q^out = {P_i^out, i = 1...n} ← Decompose Q^in same as Q^ref
initiate ε, t_ε and count_max
count ← 0
while ε ≥ t_ε do
    count ← count + 1
    ε ← 0
    for i = 1 → n do
        register P_i^out with P_i^ref in the local Cartesian Coordinate System defined by J_i using EM-ICP.
        project the transformation matrix to a 2 DOF subspace where the rotation angles are θ_i and δ_i
        ε ← ε + |θ_i| + |δ_i|
    end for
    Register Q^out with Q^ref as a whole by EM-ICP
    if count ≥ count_max then
        display('Maximum iteration reached!')
        return Q^out = {P_i^out, i = 1...n}
    end if
end while
return Q^out = {P_i^out, i = 1...n}
```

4. A Top-Bottom-Top Registration Method

Iterative alignment failure. A straightforward and intuitive registration approach is to incrementally align new frame with the existing model and add this frame to refine the model if the alignment result is proved to be good enough. Iterative Closest Points (ICP) algorithm [30] has been well developed over these years. Its state-of-the-art variant EM-ICP [21, 5] implemented on GPU [24] exhibits both robustness and speed.

In our case, articulated registration (Algorithm 1) based on EM-ICP is employed due to the existense of articulation between any new frame and our model. The model is initialized with the front pose and all succeeding frames are registered with the model. Surface Interpenetration Measure (SIM) [20] is used to check the consistency between any new frame and the model. A hard threshold is used to remove outliers, *i.e*. bad frames.

After several experiments, however, we find this approach fails for two main reasons: 1) Noisy nature of input data as described before. The larger the noise the harder it is for EM-ICP to register two points clouds. 2) EM-ICP algorithm tends to underestimate rotations while registering cylindrical shape objects [17], *i.e*. EM-ICP tends to shrink the model when you go along the cylindrical surface.

The failure of the straightforward approach enables us to think from a different perspective. In this paper, instead of registering all frames, we use 4 key poses: front (reference frame), back , left profile and right profile. We automatically detect and extract the 4 key frames out of a complete depth video sequence where a subject turns around 360^o. While the 4 frames cannot be registered by Algorithm 1 due to limited overlapping area, we register them in a top-bottom-top manner. Top-bottom means to enforce the kinematic constraints from a parent node to its child nodes while bottom-top means the opposite. For our simplified body model which consists of 5 cylinders (torso-head, left leg, right leg, left arm and right arm), top-bottom method means aligning the whole body first and then the registration of each rigid limb in the Local Cartesian Coordinate System can be restricted to a 2 DOF. The bottom-top method first aligns all rigid limbs, then the registration of torso-head is solved by considering the constraints of rigidity and connectivity. The top-bottom process can put every rigid body part in a roughly correct position while the bottom-top process refines and completes the registration.

4.1. Detect 4 Key Postures

The front pose F_{front} is labeled as soon as we detect the critical points (defined in Section 4.2). The front pose is set as the reference frame and it is used to initialize our model in the top-bottom-top registration process. Starting from the front frame we assign weighted normalized score to each succeeding frame and detect other key poses by searching over the whole sequence.

Scores associated with each frame. There are three main scores. Score s_i^m indicates the motion between two consecutive frames. Assume that the user is the only moving object in the scene, we calculate s_i^m as the sum of absolute pixel-wise difference between two consecutive range images. Score s_i^w is the width of bounding box in pixels as shown in Figure 4(a). Score s_i^s illustrates the sim-

ilarity in structure between current frame i and the front frame F_{front}. The second largest nomalized eigenvector $\hat{v_{arm}}^i$ of frame i gives us the direction of arms. Hence $s_i^s = \hat{v_{arm}}^i \cdot \hat{v_{arm}}^{front}$ indicates too what extent the current pose looks like the front pose.

Normalized scores associated with each frame. We normalize all three scores of each frame i as follows: $\bar{s_i}^m = s_i^m / \max_{i=1...n} s_i^m$, $\bar{s_i}^w = s_i^w / \max_{i=1...n} s_i^w$ and $\bar{s_i}^s = s_i^s / \max_{i=1...n} s_i^s$.

Detect the back and profiles. Based on these nomalized scores we can detect the back and two profiles.

$$s_i^{back} = \alpha s_i^m + (1-\alpha) s_i^s, \tag{3}$$

$$s_i^{profile} = \beta s_i^m + (1-\beta) s_i^w \tag{4}$$

s_i^{back} and $s_i^{profile}$ are two scores of frame i used to detect back and profiles respectively. α and β are weights for the corresponding base scores. They are all set to 0.5 in our experiment. Then the two profiles and back are detected in the following way,

$$F_{leftProfile} = \underset{i=1...n/2}{\operatorname{argmin}} s_i^{profile}, \tag{5}$$

$$F_{rightProfile} = \underset{i=n/2...n}{\operatorname{argmin}} s_i^{profile}, \tag{6}$$

$$F_{back} = \underset{i=F_{leftProfile}...F_{rightProfile}}{\operatorname{argmin}} s_i^{back}. \tag{7}$$

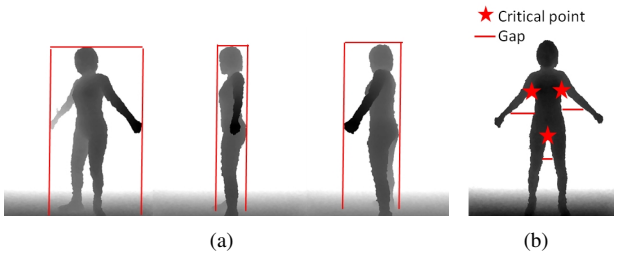
4.2. Detect Critical Points

The critical points are the points which are used for decomposing the whole body (see Figure 4(b)). Our simplified model consists of 5 cylinders, hence we need a minimum of 3 points to seperate the whole body as shown in Figure 4(b). We detect these 3 critical points from the front frame by the following procedure: 1) Detect the horizontal gaps on the front range image, *e.g.* gap between the hand and the body or gap between two legs as shown in Figure 4(b). 2) Search upwards to find the highest white pixels which are the 2D critical points. 3) Project 2D critical points to 3D critical points.

4.3. Top-bottom Registration

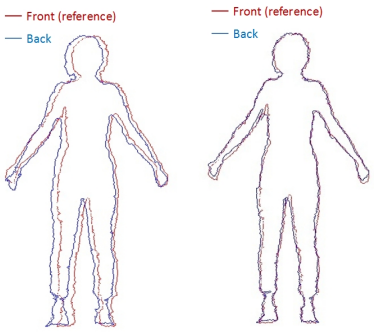
The top-bottom registration enforce the kinematic constraints from a parent to its childs. We register the back with our model using 2D silhouette information and register profiles with our model using full 3D information.

Silhouette-based top-bottom registration of back pose. First the whole back pose (*i.e.* back frame) is flipped $\varphi = \arccos(\frac{\hat{v}_{arm}^{back} \cdot \hat{v}_{arm}^{front}}{\|\hat{v}_{arm}^{back}\| \|\hat{v}_{arm}^{front}\|}) = \arccos(\hat{v}_{arm}^{back} \cdot \hat{v}_{arm}^{front})$ degrees in space along the main axis of torso and pasted to a certain depth μ. μ is predefined (*e.g.* 150mm) or roughly obtained from the two profiles. Then we register the back frame with the front frame (*i.e.* reference frame) based on their silhouettes (*i.e.* 2D points cloud) obtained from orthographic projections using Algorithm 1. The result is shown in Figure 5.



(a) (b)

Figure 4: (a) Width of the bounding box (b) Critical points detection



(a) (b)

Figure 5: (a) Before registration (b) After registration

Depth-based top-bottom registration of profile posture. The same silhouette-based method cannot be applied to register the two profile frames since the silhouette of arm cannot be found in the profile frame. We instead directly take advantage of the 3D information. After registering the back frame, we build an approximate model and fill the holes on the two sides by interpolating the cylindrical depth images. Algorithm 1 is then again used to register the two profile frames with the reference model.

4.4. Bottom-top Registration

The bottom-top registration first refines the alignment of leaf nodes and then propagate the refinement inversely along the edges until it reaches the root node. The kinematic constraint is hence enforced from the child to parent.

Refining leaf node registration. Each leaf node (besides head) represents a rigid body part, and contains information coming from 3 frames since one profile frame is occluded. The 3 points clouds are only approximately aligned after the top-bottom process and is registered more

precisely after an iterative local registration. Step a) We obtain profile's silhouette by orthographically projecting the profile points cloud along the profile direction. The correct depth and angle of back frame with respect to the front frame is then recovered from profile's silhouette. Step b) Given the front and back, we generate a cylinder initiated by front and back points clouds and fill holes by interpolation. Then we register the profile points cloud to the cylinder by EM-ICP. By iteratively executing step a) and step b), we obtain a more compact cylinder at convergence which experimentally occurs after 10 iterations. We believe that the iterative process works for two main reasons: 1) Profile points cloud contains the width information of a rigid body part. 2) Front and back points clouds have overlapping spatial feature points with the profile points cloud.

Refinement propagation. After correctly registering the child, we propagate the registration improvement to its parent. In our experiment, the parent torso-head has 4 childs, *i.e*. 4 limbs. We register the decomposed torso-head from back pose by enforcing rigidity and connectivity, *i.e*. torso-head is rigid and it is connected with four limbs. This is typically a problem of finding transformation matrix out of corresponding points which has closed-form solution.

5. Experimental Results

Figure 6 includes modeled results of 4 people scanned by our system. Each model is showed at 4 different views. Figure 6 shows clear and smoothed body shapes as a whole which contain 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.

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. We follow the decomposition idea (section 4.2) and compare the segmented rigid parts seperately. The heatmap of torso and rightleg are shown in Figure 7. 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 while the median of absolute error on right leg is 2.59mm.

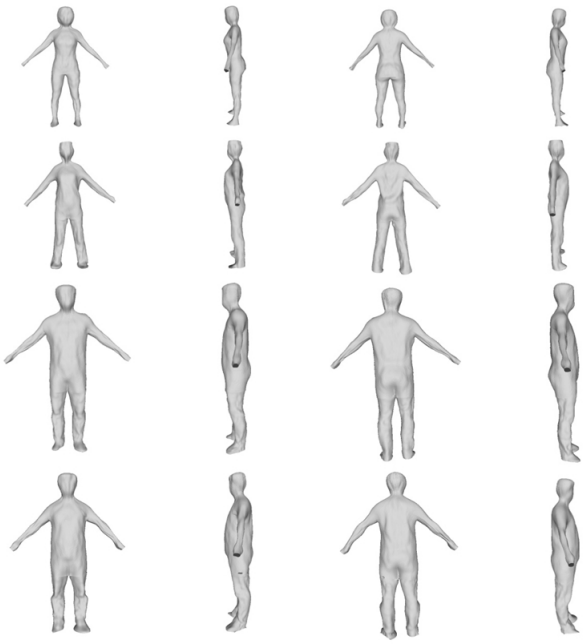


Figure 6: Different people modeled by our scanning system

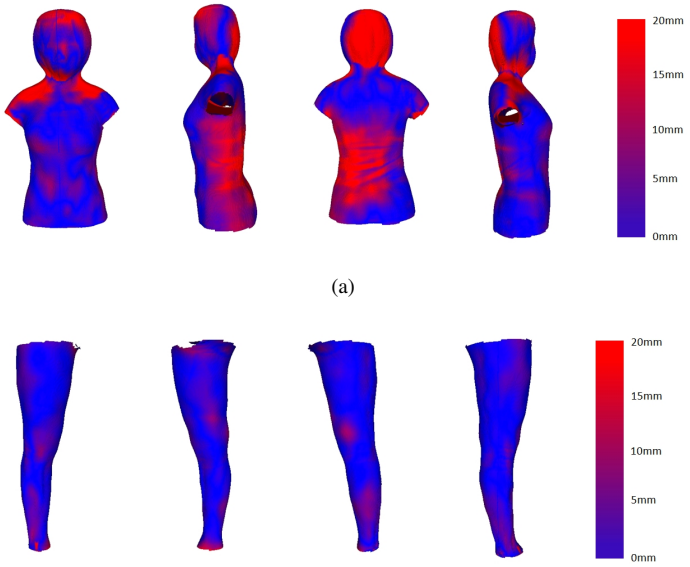


Figure 7: (a) Heatmap of torso at 4 different views (b) Heatmap of right leg at 4 different views

6. Conclusions and Future Work

We have presented a practical system for 3D body scanning that can be operated by a single naive user at home. While the accuracy result are very encouraging, we will investigate several ways to recover fine details of body

shape. These include the use of a larger number of cylinders, the exploitation of information contained in the RGB data stream and the application of super-resolution methods.

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 408–416. ACM, 2005. 3

[2] Bodymetrics. `http://www.bodymetrics.com/`. 1

[3] G. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–375. IEEE, 2003. 3

[4] K. Cheung, S. Baker, and T. Kanade. Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–77. IEEE, 2003. 2, 3

[5] S. Granger and X. Pennec. Multi-scale em-icp: A fast and robust approach for surface registration. *Computer VisionECCV 2002*, pages 69–73, 2006. 5

[6] D. Hahnel, S. Thrun, and W. Burgard. An extension of the icp algorithm for modeling nonrigid objects with mobile robots. In *International joint conference on Artificial intelligence*, volume 18, pages 915–920. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003. 3

[7] J. Hernandez, J. Choi, and G. Medioni. Laser scan quality 3-d face modeling using a low-cost depth camera. In *EUSIPCO 2012*, 2012. 4, 5

[8] Q. Huang, B. Adams, M. Wicke, and L. Guibas. Non-rigid registration under isometric deformations. In *Computer Graphics Forum*, volume 27, pages 1449–1457. Wiley Online Library, 2008. 3

[9] S. Izadi, R. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time dynamic 3d surface reconstruction and interaction. In *ACM SIGGRAPH 2011 Talks*, page 23. ACM, 2011. 3

[10] I. Jolliffe and MyiLibrary. *Principal component analysis*, volume 2. Wiley Online Library, 2002. 4

[11] K. Khoshelham. Accuracy analysis of kinect depth data. In *ISPRS Workshop Laser Scanning*, volume 38, page 1, 2011. 2

[12] Kinect. `http://www.xbox.com/en-US/kinect/`. 1

[13] K. Kolev, M. Klodt, T. Brox, and D. Cremers. Continuous global optimization in multiview 3d reconstruction. *International Journal of Computer Vision*, 84(1):80–96, 2009. 1, 3

[14] A. Laurentini. The visual hull concept for silhouette-based image understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 16(2):150–162, 1994. 1, 3

[15] H. Li, R. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. In *Computer Graphics Forum*, volume 27, pages 1421–1430. Wiley Online Library, 2008. 3

[16] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong. Modeling deformable objects from a single depth camera. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 167–174. IEEE, 2009. 3

[17] Y. Liu. Automatic registration of overlapping 3d point clouds using closest points. *Image and Vision Computing*, 24(7):762–781, 2006. 5

[18] A. Maimone and H. Fuchs. Reducing interference between multiple structured light depth sensors using motion. In *Virtual Reality Workshops (VR), 2012 IEEE*, pages 51–54. IEEE, 2012. 3

[19] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 369–374. ACM Press/Addison-Wesley Publishing Co., 2000. 3

[20] C. Queirolo, L. Silva, O. Bellon, and M. Segundo. 3d face recognition using simulated annealing and the surface interpenetration measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):206–219, 2010. 5

[21] S. Rusinkiewicz and M. Levoy. Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on*, pages 145–152. IEEE, 2001. 5

[22] Styku. `http://www.styku.com/business/benefits/`. 1

[23] R. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. In *ACM Transactions on Graphics (TOG)*, volume 26, page 80. ACM, 2007. 3

[24] T. Tamaki, M. Abe, B. Raytchev, and K. Kaneda. Softassign and em-icp on gpu. In *Networking and Computing (ICNC), 2010 First International Conference on*, pages 179–183. IEEE, 2010. 5

[25] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846. IEEE, 1998. 5

[26] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE Trans. Vis. Comput. Graph.*, 18(4):643–650, 2012. 2, 3

[27] A. Weiss, D. Hirshberg, and M. Black. Home 3d body scans from noisy image and range data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1951–1958. IEEE, 2011. 3

[28] K. Wong and R. Cipolla. Structure and motion from silhouettes. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 217–222. IEEE, 2001. 3

[29] K. Wong, P. Mendonça, and R. Cipolla. Head model acquisition from silhouettes. *Visual Form 2001*, pages 787–796, 2001. 3

[30] C. Yang and G. Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 3, 5