Quality Dynamic Human Body Modeling Using a Single Low-cost Depth Camera

Qing Zhang Bo Fu Mao Ye Ruigang Yang
University of Kentucky

Abstract

In this paper we present a novel autonomous pipeline to build a personalized parametric model (pose-driven avatar) using a single depth sensor. Our method first captures a few high-quality scans of the user rotating herself at multiple poses from different views. We fit each incomplete scan using template fitting techniques with a generic human template, and register all scans to every pose using global consistency constraints. After registration, these watertight models with different poses are used to train a parametric model in a fashion similar to the SCAPE method. Once the parametric model is built, it can be used as an animitable avatar or more interestingly synthesizing dynamic 3D models from single-view depth videos. Experimental results demonstrate the effectiveness of our system to produce dynamic models.

1. Introduction

Human body modeling, because of its many applications, has been an active research topic in both computer vision and computer graphics for a long time. In particular, with the recent availability of low cost commodity depth sensors such as the Microsoft Kinect sensor, getting the raw 3D measurement has been easier than ever. A number of approaches have been developed to make 3D models using these sensors. However, due to the relatively low-quality depths they produce, multiple overlapping depth maps have to be fused together to not only provide more coverage, but also reduce the noise and outliers in the raw depth maps. Therefore these modeling approaches are limited to static objects (e.g., the well-received KinectFusion system [9]), or human in mostly static poses (e.g, the home body scanning system [18] and the 3D self-portrait system [12]).

In this paper we present a complete system that can significantly improve the 3D model quality for human subject with dynamic motion. Our main idea is to first create a *drivable and detailed* human model, and then use the *personalized* model to synthesize a full 3D model that best fit the raw input depth map containing dynamic human motion.

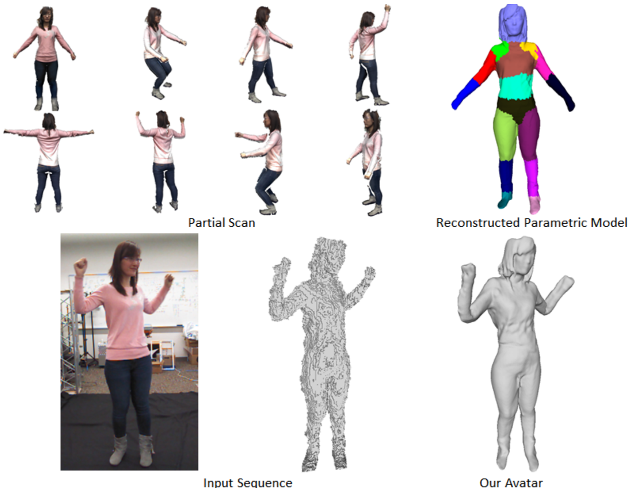


Figure 1. The pipeline of our system. We take several Kinect Fusion partial scans of different poses as initial setup (upper left) and register them to each pose. Watertight 3D models reconstructed at each pose are then used to train the pose parametric model (upper right). For an incoming video sequence, our model is drivable to fit the partial data, leading to a high-quality output model.

Our system first capture the human subject under different poses. The subject needs to stand still for a few seconds per pose while a single depth sensor that is mounted on a motorized tilt-unit scans the subject to obtain a relatively high-quality partial 3D model. Unlike previous methods, the subject does not need to rotate around and be scanned in the same pose from multiple angles. From the collection of partial scans of different poses (some from the front, some from the back, and some from the side), a *complete* 3D model is reconstructed using non-rigid point registration and merging algorithms. The model is not only personalized to the subject, but also *rigged* to support animation. Now our system is ready to synthesize high-quality dynamic models using the low-quality depth input directly from the sensors. Note that we are not simply driving the personalized model using standard skeleton-based animation techniques. In each frame, the personalized model is updated to produce a best fit to the input for the visible part. Figure 1 shows a complete example of our system. It should be noted that we achieve all of these using no more than a single depth sensor.

To the best of our knowledge, our system is the first that can automatically reconstruct a human model that is not only detailed but also drivable while using only a single commodity depth camera. Our method does not rely on any training database, requires very little user cooperation (each pose is scanned only once), and can create high-quality dynamic models of human motions. Therefore we believe our system can be used to expand the applications of depth sensors to the dynamic human modeling area.

2. Related Work

We review the related recent works in 3D human model reconstruction, mesh deformation and registration.

SCAPE-base Methods The SCAPE (Shape Completion and Animation for PEople) [1] provides a data-driven method for building the 3D human model that spans variation in both subject shape and pose and fitting the trained model to incomplete point cloud data for different body shapes in different poses.

The succedent home 3D body scan [19] applies the SCAPE approach to Kinect point cloud data and utilizes the silhouette constraints to make the fitting more robust to side views. The TenBo (Tensor-Based Human Body Modeling) [4] decomposes the shape parameters and combines the pose and shape in a tensor way to add shape variations for each body part.

Based on the general SCAPE model, many variant applications have been developed. The Naked Truth [2] estimates human body shape under clothing. DRAPE (DRessing Any PErson) [8] uses the naked body under clothing and learn the clothing shape and pose model. All these methods rely on a large training database. These result models lack facial details, hairs, and clothing effect.

Mesh Deformation and Registration The mesh embedded deformation [16] uses a rough guided graph to deform the mesh as rigid as possible. Based on the embedded model, the approach of Li *et al.* [11] uses a pre-scanned object as shape prior and register . Despite of the nonlinear embedded approach, linear mesh deformation methods such as [15, 20] are more likely to deal with small deformation and details transfer.

For handling the loop closure problem, the real time method [18] diffuses the registration error and online updates the model. This method aims to align scans of static objects. The global registration for articulated models [3] can cope with large input deformation, but is less suitable for aligning human body and garment.

The full body multiple Kinect scanning system [17] captures a dense sequence of partial meshes while the subject standing still on a turntable. All the partial scans are registered together based on the error distribution approach [14]. 3D Self-Protraits [12] presents the first autonomous capture system for self-portraits modeling using a single Kinect. The user stands as still as possible during capture and turn roughly 45 degrees at each scan.

For registering dynamic input scans without large rotation change, the global linear approarch [13] registers all the scans using the linear deformation model which assumes small rotation angle of input scans.

3. Building Complete 3D Models

In this section, we build complete 3D models for all the captured poses using partial scans. First, we introduce our data capture setup and the initial alignment using a general template model. Then we formulate the nonrigid registration problem using the embedded model of a simple yet efficient loop constraints.

3.1. Capturing System Setup

We utilize the Kinect Fusion Explorer [9] tool in Microsoft Kinect SDK to capture partial 3D meshes and colors. The subject person stands in front of the sensor approximately one meter away. The Kinect sensor is tilt from 13 degree to -27 degree during each capture. It takes four seconds per scan and the subject person keeps almost still at each pose. In order to build complete models, we take multiple scans at different angles to ensure most of body can be seen at least once.

Input meshes of Kinect Fusion are extracted from a volume of size 512^3 and 768 voxels per meter. We uniformly sample the input mesh to an average edge length of $4mm$ and erode from its boundary by $2cm$ to cut off sensor outliers. The floor is removed using background subtraction.

3.2. Template Fitting

Since there is neither a semantic information from the scanned meshes nor natural correspondences, we adopt state-of-art articulated template fitting algorithm to align a generic template onto each of the scanned inputs. Specifically, we utilize the algorithm developed by Gall et al. [7, 6]. Using a generic rigged template mesh model, this algorithm estimates the global transformation as well as joint angles of the underlying skeleton to fit the template to the input meshes. In order to better handle the single view data, we build point correspondences in a boundary-to-boundary and inner-to-inner fashion, according to the 2D projection of the meshes. Upon extraction of point correspondences, the pose is optimized iteratively via exponential map parametrization of the joint transformations. Our fitting process always starts with a standard T-pose. With the prior knowledge of rough global orientation (the subject is normally scanned in a loop fashion), the method generally provides reasonable fitting results.

3.3. Pairwise Nonrigid ICP

For pairwise registration of partial scans, we employ the embedded deformation model [16, 11], which describes plastic deformation and is effective to handle articular human motion [11]. The embedded method defines the deformation of each vertex $\mathbf{v}$ on the mesh influenced by m nearest nodes $\mathbf{g}$ on a coarse guide graph. In our case, two meshes M_i, M_j have already aligned with their graphs T_i, T_j after our template fitting step, and also T_i, T_j have the same face connectivity. The transformation from T_i to T_j is defined on each node $\mathbf{g}^k$: a 3×3 affine matrix $\mathbf{R}_i^k$ and a translation vector $\mathbf{t}_i^k$. Given transformations, the node on deformed graph $\tilde{T}_i$ is simply added the translation: $\tilde{\mathbf{g}}^k = \mathbf{g}^k + \mathbf{t}_i^k$ on the graph and the deformed vertex is computed as $\tilde{\mathbf{v}} = \sum\limits_{l=1}^{m} w_l(\mathbf{v}_i)\left[\mathbf{R}_l(\mathbf{v}_i - \mathbf{g}_l) + \mathbf{g}_l + \mathbf{t}_l\right]$ where $w_l(\mathbf{v}_i)$ is the influence weight inversely proportional to the distance from $\mathbf{v}_i$ to its control nodes $\|\mathbf{v}_i - \mathbf{g}_l\|$.

It can be easily verified that if $\left(\mathbf{R}_1^k, \mathbf{t}_1^k\right)$, $\left(\mathbf{R}_2^k, \mathbf{t}_2^k\right)$ are two consecutive deformations of T_i, the total deformation is $\left(\mathbf{R}_2^k\mathbf{R}_1^k, \mathbf{t}_2^k\mathbf{t}_1^k\right)$. Let $\mathbf{R}_2^k = \left(\mathbf{R}_1^k\right)^{-1}$ and $\mathbf{t}_2^k = -\mathbf{t}_1^k$, then the mesh deformed by $\left(\mathbf{R}^k, \mathbf{t}^k\right)$ can be restored using $\left(\left(\mathbf{R}^k\right)^{-1}, -\mathbf{t}^k\right)$. We assume all the $\{\mathbf{R}^k\}$ are almost rigid and this property holds in our case.

For registering M_i to M_j, transformations $\left(\mathbf{R}_i^k, \mathbf{t}_i^k\right)$ are solved by minimizing the energy function [11]

$$\min\left(w_{rot}E_{rot} + w_{reg}E_{reg} + w_{fit}E_{fit}\right), \qquad (1)$$

where $E_{rot} = \sum_k Rot(\mathbf{R}_i^k)$, $Rot(\mathbf{R})$ specifies the orthogonality and rigidity of the transformation. $E_{reg} = \sum\limits_{k}\sum\limits_{l \in N(k)} \left\|\mathbf{R}_i^k(\mathbf{g}_i^l - \mathbf{g}_i^k) - (\mathbf{g}_i^l + \mathbf{t}_i^l - \mathbf{g}_i^k - \mathbf{t}_i^k)\right\|^2$ ensures smoothness of the deformation.

The fitting term $E_{fit} = \sum_c \alpha_{point}\|\mathbf{v}_i^c - \tilde{\mathbf{v}}_i^c\|^2 + \alpha_{plane}\left|\tilde{\mathbf{n}}_i^T(\mathbf{v}_i^c - \tilde{\mathbf{v}}_i^c)\right|^2$ constrains the deformed position of a subset of vertices, where $\tilde{\mathbf{v}}_i^c$ specifies the destination of $\mathbf{v}_i^c$ and $\tilde{\mathbf{n}}_i$ is the normal on the surface of M_j accordingly. Different from [11], since associated T_i and T_j have the same face, we are able to segment each mesh by corresponding graph nodes as shown in Figure 2. The same colored region denotes vertices influenced by same graph nodes. When searching correspondences from M_i to M_j, we perform iterative closest point (ICP) algorithm to align large patches (area > threshold) and search for the closest point after ICP. Faraway or normal inconsistent pairs are excluded. We obtain in roughly 2000 correspondences for a pair of scans.

The cost function equation 1 is minimized by Gauss-Newton solver and see [16, 11] for details. After registration, we get all of the transformations $\left\{\left(\mathbf{R}_i^k, \mathbf{t}_i^k\right)\right\}$, the deformed graph, the deformed mesh and a corresponding point set. We set a large rigid weight w_{rot} to maintain high stiffness during a sequence of deformations. Another tradeoff is to set a relatively larger regularization weight w_{reg} and smaller fitting weight w_{fit}. It results in slower convergence to correct destination of the overall algorithm but benefits the avoidance of severe failure deformation of the graph such as self intersection and volume collapse due to error accumulation. In our experiment, it shows that the whole algorithm still converges within 5-10 iterations as Figure 5.

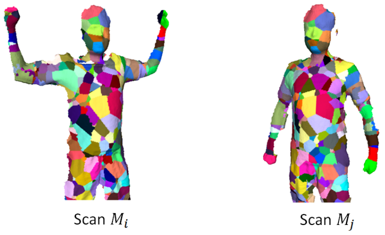


Figure 2. We search for corresponding points by aligning patches controlled by the same graph nodes using ICP.

3.4. Global Nonrigid Registration Algorithm

We have n partial scans in the capture step 3.1 and they are aligned with graphs in 3.2. In this section,we register all scans to each pose while achieving global geometry consistency. Inspired by [14, 17], we develop an iterative optimization scheme to 1) pairwise register scans and 2) adjust them by distributing accumulative error using loop closure constraints. Different from the method in [17], since the deformation of a graph is simply adding the translation $\mathbf{t}_i^k$ to each node, $\mathbf{R}_i^k$ does not interfere with the graph directly. Therefore, we deal with translational and rotational error distribution separately, and translational error optimization is simpler and more efficient.

Preprocessing Given input scans and graphs, we initially register all the graphs to the target graph and deform all scans accordingly as shown in Figure 3. To suppress outliers occurring near joints, we remove faces of long edge length and clean disconnected small patches from the deformed mesh. To reduce the influence of badly deformed vertices, we compute the affine transformation near each vertex and compare the deviation angle of the corresponding Laplacian coordinates. Each vertex is assigned to a confidence weight W_{lap} inversely proportional to the deviation.

After the rough registration, the covered region on the target graph of each scan is known. By aligning the torso part (chest and abdomen), we can roughly determine each virtual camera pose in the target coordinate system. Sorting angles from the target camera to each virtual camera, we finally get a circle of n scans denoted as $M_1, M_2, \ldots, M_n$ and the target scan $w.l.o.g.$, is denoted as M_1 in Figure 3.

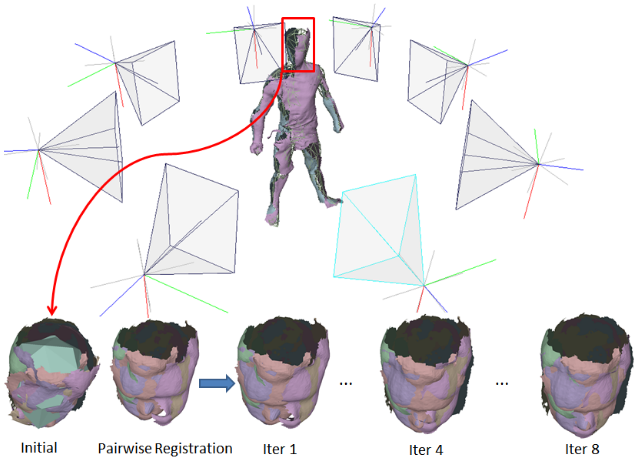


Figure 3. Stages in our global registration. All the partial scans are initially aligned to the target using the general template model. Virtual cameras are estimated in the coordinate system of the target pose to determine the loop closure. The fitted template model is reduced to a rough graph to guide the embedded registration. Pairwise accumulated registraton error is distrubuted after each loop adjustment.

Bi-directional Loop Constraints Now we have a loop of n scans $M_i, i = 1, \dots, n$, the graph T_1 is aligned with M_1 correctly and we use it as the embedded graph to register M_1 to M_2 by using the deformation described in 3.3. After the registration, M_1, T_1 are deformed as $M_{1,2}, T_{1,2}$ and transformations are denoted as $\left\{\left(\mathbf{R}_1^k, \mathbf{t}_1^k\right)\right\}$. By using the weight and node indices of T_2 but the node positions of $T_{1,2}$, we register M_2 to M_3 and get $M_{2,3}, T_{2,3}$. The process continues until registering T_n back to T_1 with transformations $\left\{\left(\mathbf{R}_n^k, \mathbf{t}_n^k\right)\right\}$. We call this step as the pairwise registration in the context of this section. For a globally correct registration, we have $T_{n,1} = T_1$, that is for each node, $\mathbf{t}_1^k + \mathbf{t}_2^k + \cdots + \mathbf{t}_n^k = 0$, and the deformed mesh $M_{n,1}$ is consistent with M_1. When the deformation is highly rigid, applying the multiplication of consecutive deformations, the product of rotations along the loop will be an identity, that is $\mathbf{R}_n^k \mathbf{R}_{n-1}^k \cdots \mathbf{R}_1^k = \mathbf{I}$.

Due to error accumulation, the pairwise registration will drift and violate such constraints. Similar to [17], we distribute the accumulated rotational and translational error and choose a weight $w_i = 1/Dist(M_{i,i+1}, M_{i+1})$ to transformations $\left\{\left(\mathbf{R}_i^k, \mathbf{t}_i^k\right)\right\}$, where $Dist(M_{i,i+1}, M_{i+1})$ is the average fitting error of E_{fit} in 1, for all $i = 1, \dots, n$. ($n+1$ we refer to 1.) Since every node will be optimized in the same way, we ignore the superscript k in the following.

The translational error is distributed by solving the following optimization,

$$\min \sum_{i=1}^{n} w_i^2 \left\|\hat{\mathbf{t}}_i - \mathbf{t}_i\right\|^2, \quad s.t., \sum_{i=1}^{n} \mathbf{t}_i = 0, \qquad (2)$$

and the solution is found using Lagrange multipliers, $\hat{\mathbf{t}}_i = \mathbf{t}_i - \alpha_i \sum_{j=1}^{n} \mathbf{t}_j$, with the scalar α_i as

$$\alpha_i = \frac{1}{w_i^2} \Bigg/ \sum_{j=1}^{n} \frac{1}{w_j^2} \qquad (3)$$

The rotational error distribution is to minimize the total rotational deviation:

$$\min \sum_{i=1}^{n} w_i \angle(\hat{\mathbf{R}}_i, \mathbf{R}_i), \quad s.t., \mathbf{R}_n^k \mathbf{R}_{n-1}^k \cdots \mathbf{R}_1^k = \mathbf{I}, \quad (4)$$

where the angle between two rotations is defined as $\angle(\mathbf{A}, \mathbf{B}) = \cos^{-1}\left(\frac{tr(\mathbf{A}^{-1}\mathbf{B})-1}{2}\right)$. Analyzed in [14], the optimal $\hat{\mathbf{R}}_i$ is computed as

$$\begin{aligned} \hat{\mathbf{R}}_i &= \mathbf{E}_i^{<\alpha_i>} \mathbf{R}_i, \\ \mathbf{E}_i &= (\mathbf{R}_k \mathbf{R}_{k-1} \cdots \mathbf{R}_1 \mathbf{R}_n \mathbf{R}_{n-1} \cdots \mathbf{R}_{k+1})^{-1}, \end{aligned} \qquad (5)$$

where α_i is referred to equation 3, and $\mathbf{E}_i^{<\alpha_i>}$ is defined to be the rotation matrix that shares the same axis of rotation as $\mathbf{E}_i$ but the angle of rotation has been scaled by α_i.

Once all the optimal $\left\{\left(\hat{\mathbf{R}}_i^k, \hat{\mathbf{t}}_i^k\right)\right\}$ are obtained, we use the total transformation $\left\{\left(\left(\hat{\mathbf{R}}_1^k \dots \hat{\mathbf{R}}_{i-1}^k \hat{\mathbf{R}}_i^k\right)^{-1}, -\hat{\mathbf{t}}_i^k - \hat{\mathbf{t}}_{i-1}^k - \cdots - \hat{\mathbf{t}}_1^k\right)\right\}$ to deform the mesh M_i with $T_{i-1,i}$ back to M_1. After all the meshes M_i updated, we repeat the pairwise registration step from M_1 and T_1. The graphs $T_1, T_{1,2}, \dots, T_{n,1}$ will finally converge to a constant graph and $\left\{\left(\hat{\mathbf{R}}_i^k, \hat{\mathbf{t}}_i^k\right)\right\}$ converges to the globally optimal solution as plotted in Figure 5.

In the sense that the error distribution step can prevent graph drifting and pull it towards the optimal position, we can perform an interleaved bi-directional way to avoid large accumulative errors. The basic idea is to perform an inverted iteration using the order of $M_1, M_n, M_{n-1}, \dots, M_3, M_2, M_1$ after a forward directional iteration. The directional scheme is in essential the same to the multiple cycle blending technique described in [14] and the total time complexity to convergence is the same because they traverse in both direction in one iteration and we perform in each direction once but need two iterations.

3.5. Postprocessing

Once all the partial scans are registered to the target pose, the final water-tight surface is extracted by using Screened Poisson Surface method [10] which takes the point confidence into account. We assign a blending confidence for each point $W = W_{normal} * W_{sensor} * W_{lap}$: W_{normal} is inversely proportional to the angle between the original input normal and the z-axis; W_{sensor} is proportional to the distance from a point to the mesh boundary; W_{lap} is inversely proportional to the deviation Laplacian coordinates, and the final weight W is pruned to $[\epsilon, 1]$, $\epsilon > 0$. The surface color is transferred from the input color and diffused using Poisson blending method [5] to achieve seamless.

4. Training the Personalized Model

In this section, we align the example 3D models built in the above section to train the animatable parametric model and fit it to the new incoming depth sequence. Different from the SCAPE based methods [1, 19, 4], which varies at the ability of representing personal details. Our complete models are inherently specified to a certain user and have no shape variations, therefore we only need to train the regression vectors of joints for a personalized model.

Before training the parametric model, all of the 3D models are required to be mesh topology consistent. We pick a neutral pose as the reference pose and deform it to all the other 3D models. Similar to the nonrigid ICP registration in section 3.3, we register the reference model to each complete model by taking the alignment of their associated graphs as the initial guess. As a result of nonrigid ICP registration, corresponding points are found with normal consistency. We employ the detail synthesize method to make subtle adjustment of the warped reference model:

$$\min_{d_i} \sum_{\mathbf{v}_i} \|\mathbf{v}_i + d_i \mathbf{n}_i - \mathbf{v}_i^c\|^2 + \beta \sum_{i,k} |d_i - d_k|^2, \qquad (6)$$

in which $\mathbf{v}_i$ and $\mathbf{v}_i^c$ are corresponding points, d_i is the distance along its normal direction $\mathbf{n}_i$. The distance field is diffused among neighboring vertices i and k. $\beta = 0.5$ in the experiments.

After registered to all the other $n-1$ example poses, the reference model is ready for training. First, we transfer the body part index (16 parts in total) from the generic body template 3.2 to the reference model using nearest neighboring searching as shown in Figure 1. Then considering sample i for each body part l, a rigid rotation $\mathbf{R}_l^i$ is solved using ICP. For each face k of part l, a 3×3 nonrigid transformation matrix $\mathbf{Q}_k^i$ is solved via the following equation:

$$\min_{\mathbf{Q}_k} \sum_k \sum_{j=2,3} \left\|\mathbf{R}_k^i \mathbf{Q}_k^i \hat{\mathbf{u}}_{k,j} - \mathbf{u}_{k,j}^i\right\|^2 + \rho \sum_{k_1,k_2} \left\|\mathbf{Q}_{k_1}^i - \mathbf{Q}_{k_2}^i\right\|^2, \qquad (7)$$

where k_1, k_2 are neighboring faces, $\mathbf{u}_{k,j} = \mathbf{v}_{k,j} - \mathbf{v}_{k,1}$, $j = 2, 3$ are two edges, $\rho = 1e^{-3}$ is to prevent the large deformation change.

Given all the $\mathbf{Q}_k^i$ and joint angles computed from $\mathbf{R}_k^i$, a regression matrix $\mathbf{A}$ mapping joint angles to $\mathbf{Q}_k^i$ can be trained from samples similar to SCAPE method [1]. Note that in our case, we have less (usually $n = 8$) poses than the SCAPE training data. However, since the regression is linear, the ability of its representation depends on the range of joint angles instead of number of samples. In our capture stage, the subject person is required to perform different joint configurations as much as possible. And then the trained model ends up being able to recover the personalized style of movement.

The trained model allows us to fit new incoming point cloud in an ICP scheme, which is formulated as an optimization to solve $\mathbf{R}_k$ of each body part given point correspondences. Specially, it can be formulated as an optimization problem to minimize the energy:

$$\begin{aligned} \min_{\mathbf{R}_k} \sum_k \sum_{j=2,3} & \|\mathbf{R}_k \mathbf{Q}_k \hat{\mathbf{u}}_{k,j} - \Delta \mathbf{y}_{k,j}\|^2 \\ + w_p \sum_m & \|\mathbf{y}_m - \mathbf{y}_m^c\|^2 \end{aligned} \qquad (8)$$

where $\mathbf{y}_k$ are the vertices on the reconstructed mesh, $\{\mathbf{y}_m\}$ and $\mathbf{y}_m^c$ are corresponding points and $w_p = 1$ is a weight to balancing the influence of correspondences. Since all the $\mathbf{R}_k$, $\mathbf{Q}_k$, $\mathbf{y}_k$ are unknown, the optimization is nonlinear. Using the similar technique as [1], given an initial guess of $\mathbf{R}_k$, the other two terms $\mathbf{Q}_k$ and $\mathbf{y}_k$ can be solved in a linear least square accordingly. Once $\mathbf{Q}_k$ and $\mathbf{y}_k$ are given, the rotation $\mathbf{R}_k$ can be updated again by a twist vector ω, $\mathbf{R}_k^{new} \leftarrow (\mathbf{I} + [\omega]_\times)\mathbf{R}_k^{old}$, in which $[\cdot]_\times$ denotes the cross product matrix. The twist vector ω is then solved by minimizing:

$$\begin{aligned} \min_{\omega_k} \sum_k \sum_{j=2,3} & \|(\mathbf{I} + [\omega]_\times)\mathbf{R}_k \mathbf{Q}_k \hat{\mathbf{u}}_{k,j} - \Delta \mathbf{y}_{k,j}\|^2 \\ + w_t \sum_{l_1,l_2} & \|\omega_{l_1} - \omega_{l_2}\|^2, \end{aligned} \qquad (9)$$

in which l_1 and l_2 denote two neighbor body parts. It is a linear least square problem and can be solved efficiently for 16×3 unknowns in total. After alternatively updating $\mathbf{R}_k$, $\mathbf{Q}_k$ and $\mathbf{y}_k$ until converging to a local minima, point correspondences are updated to the newly fitted model by searching closet points. The total ICP scheme repeats until reaching a maximum number of iterations.

5. Results

We validated our system by scanning the mannequin for performance evaluation and accuracy comparison. We scanned male and female subjects at several challenging poses to build 3D model training samples. We captured several video sequences to validate the fitting using our trained model.

Mannequin Validation As an accuracy test of our system pipeline, we acquired a 3D model of an articular mannequin and compared our results to a model captured using a high-performance structured light scanner with a 0.5mm spatial resolution.

In this test, we manually turned the mannequin around by approximately 45 degrees at each time. The mannequin was not totally rigid, and its arms and legs were slightly moved when turned around. In this case, we directly perform the pairwise registration step with loop closure adjustment. We compare it with the groundtruth to achieve an average alignment error of 2.45mm. We also compare the result with the previous paper [12] and the comparable result is shown in Figure 4.

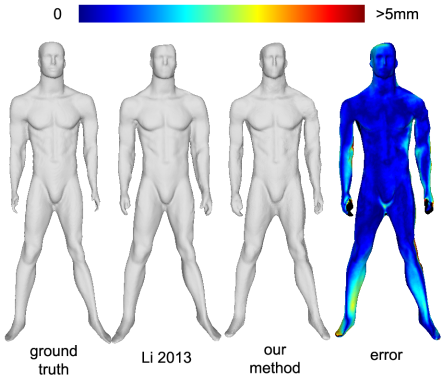


Figure 4. The reconstructed mannequin of an almost static pose. Error map compared to the groundtruth is plotted.

In another mannequin set, we test the performance of our system by capturing large pose changes. The mannequin's arms and legs were articulately moved to several poses. The qualitative evaluation results are shown in Figure 6. In Figure 5, we show the algorithm performance to register all scans to the target pose 3.4. According to the results, the optimization procedure converges in $5 - 10$ iterations for both rotational and transnational error distributions. The final average variation in rotation is less than 0.5 degree and the variation in translation is less than $0.1mm$, which we set as a terminating condition for real person modeling.

Real Person Examples We validate our system to reconstruct both female and male persons in regular clothes. It takes several minutes to capture static scans and then watertight example poses are reconstructed as shown in Figure 7. We pick the neutral pose as the reference and train parametric model. The final avatar is at the resolution of 100k faces. Figure 8 shows the fitting error.

Driving and Fitting to Video Sequence After training the parametric model, we test our drivable avatar using the full body video sequence at a distance about $2m$ to the Kinect sensor. Our parametric model is initially driven to the pose estimated by skeleton and then iteratively fitted to the input point cloud. Figure 9 shows several frames of our final fitting result. See the supplementary video for both result sequences.

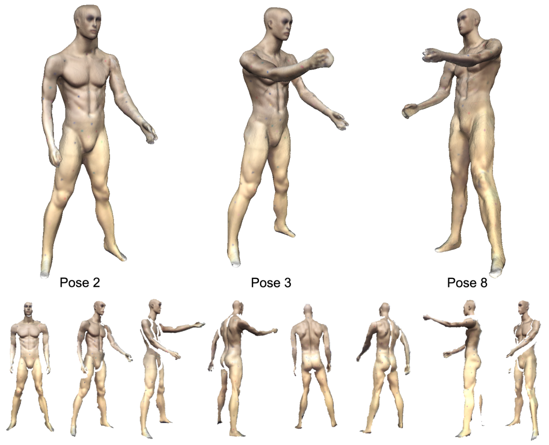
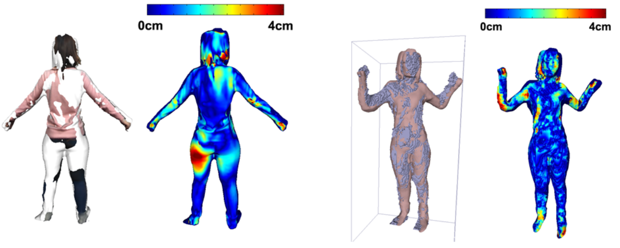


Figure 6. The reconstructed mannequin of some articulated arm movement.



(a) Final fitting to input scan 19 and the error map

(b) Final fitting to low resolultion Kinect sequence frame 105 and the error map

Figure 8. The fitting error from the reference model to input Kinect Fusion scan and input depth sequence.

Limitations Our registration method has limited power to handle highly nonrigid deformations such as loose garments. The unsmooth texture colors are mainly affected by shadow of wrinkles. Our avatar model does not model the human expression either.

6. Conclusion

We present in this paper an automatic system to create dynamic human models from a single low-quality depth camera. Our system first captures the human subject under different static poses using multi-frame fusing techniques. From the collection of partial but high-quality scans, a *complete* 3D model is reconstructed using non-rigid point registration and merging algorithms. The model is not only personalized to the subject, but also *rigged* to support animation. With that personalized avatar, our system is ready to synthesize high-quality dynamic models using the low-quality depth input directly from the sensors.

With home application in mind, our system requires minimum hardware requirement and is in particular user-friendly: the static poses are only scanned once. We have extended a few state-of-the-art point processing and model-

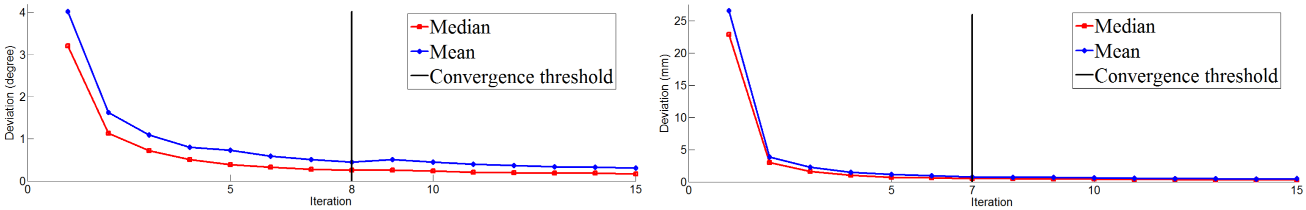


Figure 5. The deviation of mannequin data. The left is the rotation angle changes in degrees and the right is the translation in milimeters.

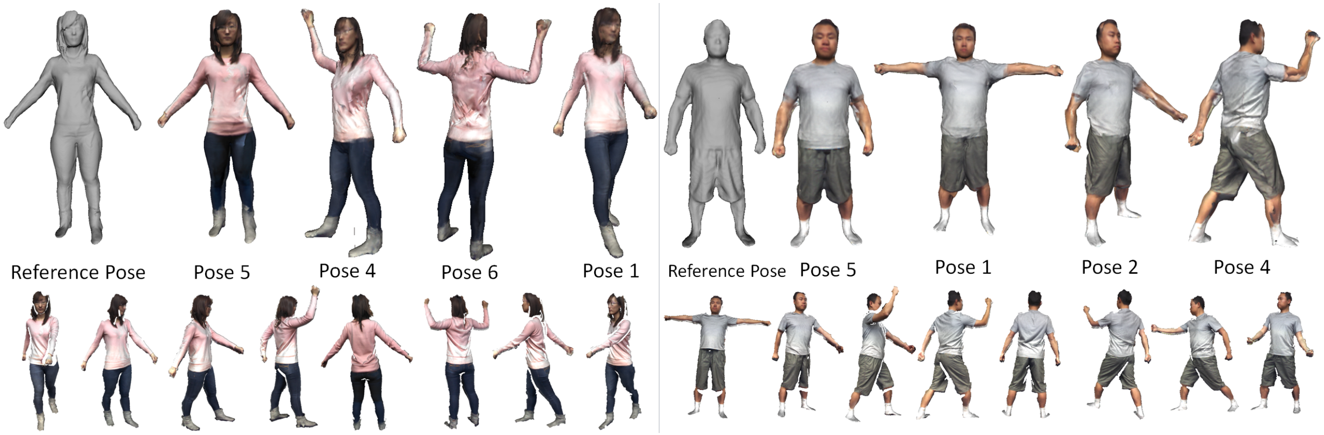


Figure 7. The reconstructed watertight models after our global registration. The bottom row shows the input partial scans and the upper row shows the reconstructed models at each pose.

ing algorithms to intelligently merge partial scans with large variations of poses to form a complete rigged model. Using only a single commodity depth camera, our approach generates dynamic avatar models with significantly more details than existing state-of-the-art human modeling approaches. Therefore it can have broad applications in simulation and training, gaming, and 3D printing, in which human modeling is a crucial part.

Acknowledgement This work is supported in part by US National Science Foundation grants IIS-1231545 and IIS-1208420, and Natural Science Foundation of China (No.61173105,61373085, 61332017).

References

[1] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. Scape: shape completion and animation of people. In *SIGGRAPH*, pages 408–416, 2005.

[2] A. Blan and M. Black. The naked truth: Estimating body shape under clothing. In D. Forsyth, P. Torr, and A. Zisserman, editors, *ECCV*, pages 15–29. 2008.

[3] W. Chang and M. Zwicker. Global registration of dynamic range scans for articulated model reconstruction. *ACM Trans. Graph*, 30(3), 2011.

[4] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In *CVPR*, pages 105–112, 2013.

[5] M. Chuang, L. Luo, B. J. Brown, S. Rusinkiewicz, and M. Kazhdan. Estimating the Laplace-Beltrami operator by restricting 3d functions. *Symposium on Geometry Processing*, July 2009.

[6] J. Gall, A. Fossati, and L. Van Gool. Functional categorization of objects using real-time markerless motion capture. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1969–1976, 2011.

[7] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H. P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1746–1753, 2009.

[8] P. Guan, D. Reiss, L.and Hirshberg, A. Weiss, and M. J. Black. Drape: Dressnig any person. *SIGGRAPH*, 2012.

[9] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, UIST '11, pages 559–568. ACM, 2011.

[10] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29:1–29:13, July 2013.

[11] H. Li, B. Adams, L. J. Guibas, and M. Pauly. Robust single-view geometry and motion reconstruction. In *SIGGRAPH Asia*, pages 175:1–175:10, 2009.

[12] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *SIGGRAPH Asia*, 32(6), November 2013.

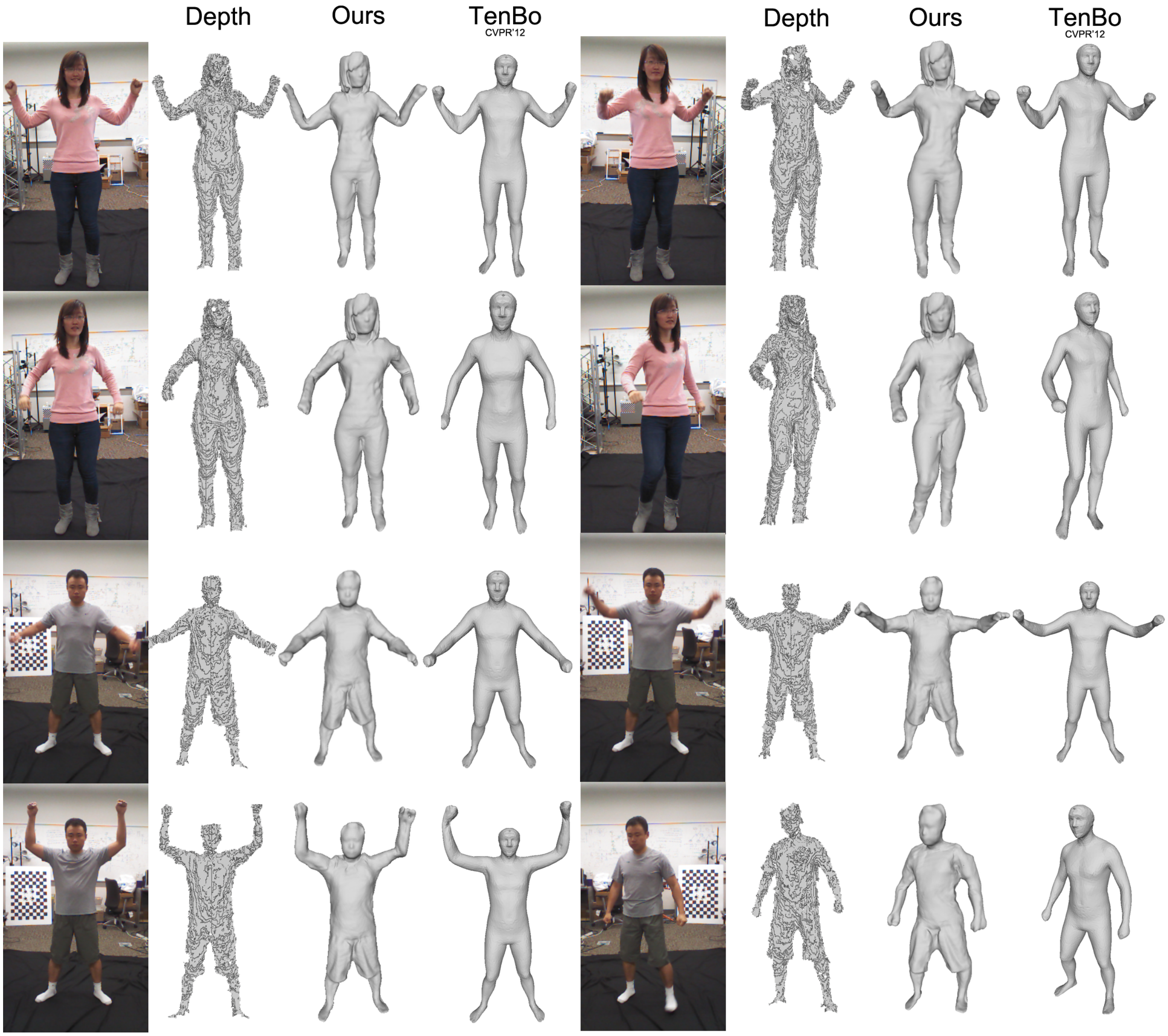


Figure 9. The final fitting result with our personalized parametric avatar. We compare our avatar with the general SCAPE model to show more realistic details.

[13] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong. Modeling deformable objects from a single depth camera. In *ICCV*, pages 167–174, 2009.

[14] G. Sharp, S.-W. Lee, and D. Wehe. Multiview registration of 3d scenes by minimizing error between coordinate frames. *PAMI*, 26(8):1037–1050, 2004.

[15] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, SGP '04, pages 175–184, New York, NY, USA, 2004. ACM.

[16] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Trans. Graph.*, 26(3), July 2007.

[17] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *TVCG*, 18(4):643–650, 2012.

[18] T. Weise, T. Wismer, B. Leibe, and L. J. V. Gool. Online loop closure for real-time interactive 3d scanning. *Computer Vision and Image Understanding*, 115(5):635–648, 2011.

[19] A. Weiss, D. Hirshberg, and M. Black. Home 3d body scans from noisy image and range data. In *ICCV*, pages 1951–1958, 2011.

[20] L. Zhang, N. Snavely, B. Curless, and S. M. Seitz. Spacetime faces: High-resolution capture for modeling and animation. In *ACM Annual Conference on Computer Graphics*, pages 548–558, August 2004.