Journal of Manufacturing Systems 33 (2014) 233–240

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Technical Paper

Fast capture of personalized avatar using two Kinects

Yin Chen [a], Gang Dang [a,*], Zhi-Quan Cheng [b], Kai Xu [a,*]

[a] PDL Laboratory, School of Computer, National University of Defense Technology, Changsha City, Hunan Province 410073, China
[b] Avatar Science Company, China

ARTICLE INFO

Article history:
Received 11 August 2013
Received in revised form
11 November 2013
Accepted 14 November 2013
Available online 11 January 2014

Keywords:
Human body capture
Kinect
Rigid alignment
Personalized avatar

ABSTRACT

We present a system for fast capture of personalized 3D avatar using two Kinects. The key feature of the system is that the capturing process can be finished in a moment, or quantitatively 3 s, which is short enough for the person being captured to hold a static pose stably and comfortably. This fast capture is achieved by using two calibrated Kinects to capture the front and back side of the person simultaneously. To alleviate the view angle limit, the two Kinects are driven by their automatic motors to capture three scans covering the upper, middle and lower part of the person from front and back respectively, resulting in three partial scans for each Kinect. After denoising, all partial scans are rigidly aligned together using a novel supersymmetric third-order graph matching algorithm. Since all these partial scans can be captured in a moment, the discrepancy between them caused by body movement is neglectable, saving the effort of non-rigid alignment. The missing gaps between the front and back scans are filled using quadratic Bézier curve. The final reconstructed mesh model demonstrates good fidelity against the person with personalized details of hairstyle, face, and salient cloth wrinkles.

1. Introduction

The rapid growing of somatosensory interaction technique is becoming a new and strong momentum for the development of computer graphics applications. For example, the recent release of Microsoft Kinect [1] has quickly made somatic game unprecedentedly prevalent. A key component of somatosensory interaction is 3D avatar, which is a digitalized 3D representation of a user or her/his alter ego. The user can drive her/his 3D avatar to interact with the virtual world.

One important goal in building 3D avatar is to make it as similar to the user as possible, hence distinguishable from that of other one's, leading to the so-called personalized avatar. There are many factors for a personalized avatar, such as face, clothing, hairstyle, etc. Among all the factors, body shape is a good trade-off between modeling difficulty and reliability. In this work, we use Kinect to construct personalized 3D avatar for any person, offering not only accurate body shape but also moderate details of hairstyle, face, and salient cloth wrinkles.

Traditional approaches to full body human capturing and modeling often rely on complicated and expensive setup, making it difficult for a casual user to create her/his virtual clone out of laboratory. Such examples are like the 3D scanning systems of SCAPE [2] and Cyberware [3]. Alternatively, Microsoft Kinect, as a low-price depth camera, has recently been used to capture human body [4,5]. However, the depth data captured by Kinect over a certain distance is of extreme low quality, making it hard to be directly used to construct accurate 3D model. Weiss et al. [4] capture *naked* full bodies using a single Kinect. To obtain accurate 3D shape estimates, the user has to perform a serial of varying poses in front of the sensor. The multiple monocular views are combined to compute body shape and pose simultaneously using the SCAPE model [2].

In our system, we leave out the pose factor and aim at fast capture of dressed human body in a moment, which is like camera shooting. A similar system was presented by Tong et al. [5], where a 3D full body is obtained by using three calibrated Kinects to capture a stationary person standing on a turntable *for about 30 s*. In contrast, our system builds a more flexible and simple setup, i.e., two easily calibrated Kinects but without turntable. More importantly, the capture can be finished *in 3 s* and the total time for building a personalized avatar is about 1 min (compared to several minutes by Tong et al. [5]). The former feature is made possible by a simple geometric calibration method while the latter benefits from a novel fast high-order graph matching algorithm. These unique features make our system user-friendly, and especially suitable for out-of-laboratory environment.

In psychology studies, a moment, qualitatively measured as 3 s, is suggested to be the basic temporal building blocks of a person behaviorally expressing subjective experiences [6]. Perceptually, a moment is seen to be a comfortable period for a general person to perform a stationary pose. This motivates us to build a user-friendly 3D avatar capturing system where the capture process can

* Corresponding authors.
E-mail addresses: gangdang@nudt.edu.cn (G. Dang), kaixu@nudt.edu.cn (K. Xu).

0278-6125/$ – see front matter © 2013 The Society of Manufacturing Engineers. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jmsy.2013.11.005

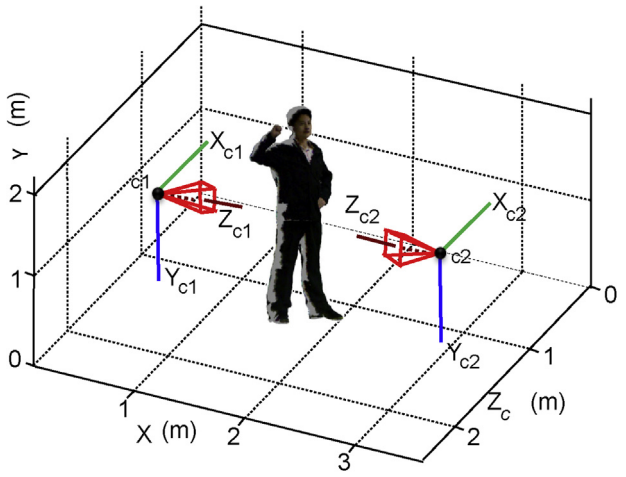


Fig. 1. Illustration of the system setup: two Kinects in opposite orientations are located at the front and back sides of the person being captured, with 1.5–3 m mutual distance and half of the human height.

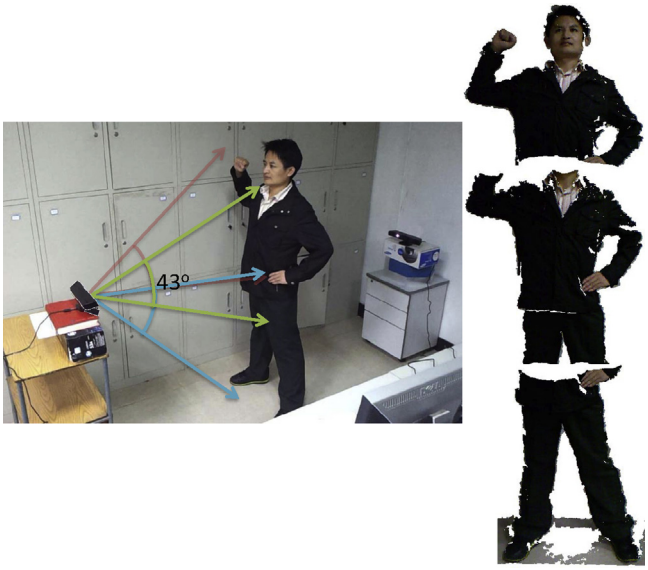


Fig. 2. A snapshot of our system (left) and three partial scans covering the upper, middle and lower part of the human body captured by the front Kinect.

be finished in a moment. To the best of our knowledge, this is the first human capturing system satisfying such time requirement. Moreover, our system can obtain qualified 3D surface models with personalized details such as hairstyles and salient cloth wrinkles (see Fig. 1).

2. Related work

The typical 3D scanning device used for capturing detailed human bodies is laser scanner, which is an active off-the-shelf device that uses laser to capture the subject. The Cyberware Whole Body Color 3D Scanner [3] can capture a personalized avatar in about 17 s. However, the system is expensive for a casual user. In addition, the user is required to hold a static pose for 17 s, which is much longer than a moment.

The Microsoft Kinect [1] is a much cheaper alternative. Many pioneer studies have been conducted on using Kinects for scanning objects. Izadi et al. [7] propose a GPU-based implementation for simultaneous Kinect camera tracking and rigid surface reconstruction. Cui et al. [8] develop a system to scan coarse 3D body of a person turning around continuously in front of a static Kinect for about 30 s. Weiss et al. [4] build a nice system to capture the 3D body of a moving person using a single Kinect, assisted by the parameterized SCAPE model [2]. Both silhouettes in RGB image and depth data are used in shape and pose estimation. Due to the limit dimension of PCA used in the SCAPE model, the quality of reconstructed model is not sufficiently high to be person-wise distinguishable. By using three Kinects and some accompanied devices, Tong et al. [5] present a successful system to capture 3D human body models in about 30 s. The applications of these existing systems are somewhat limited due to the complexity of system setup and the capturing time.

So far, none of the existing systems can fulfill the requirement of fast capture, i.e., in 3 s, which is suggested as a proper period for a general person to take a static pose stably and comfortably, according to the studies from psychology [6,9]. In this paper, this time requirement is considered as an important target in building our capturing system. This paper is an extended version of our paper appeared in SIGGRAPH ASIA Posters 2012 [10].

3. Overview

The configuration of our capturing system is illustrated in Fig. 1. Two Kinects, with opposite orientations, are located at the front and back sides of the person being captured. The distance between the two Kinects is about 1.5∼3 m. The height of both Kinect cameras is about the half height of the person. Note that this configuration is rather flexible due to the convenient calibration mechanism used in our system (see Section 4.1).

During capturing, the user stands in the middle of the two Kinects, holding a static pose. Basically, the user can take any pose that is occlusion-free, as shown in Fig. 1. Since the view angle of Kinect camera is 43° and the valid scanning distance is limited, each Kinect can only scan a part of the human body. Therefore, we use the motor driver installed on the Kinect camera to adjust its pitching angles to −30°, 0° and +30°, capturing three frames covering the upper, middle and lower parts of the human body. Thus, the two Kinects capture six frames. Each frame is a partial scan of the human body including a 640 × 480 color image registered with a 640 × 480 depth image (e.g. Fig. 2). The 3D coordinates of the scanned points are automatically computed with the OpenNI package [11]. Using OpenNI, we can also work with Kinect's motor driver and obtain self registration between color image and depth image.

4. Avatar capture and reconstruction

After the two Kinects are properly placed and oriented, our system is physically set up. Our system works in six stages: Kinect calibration, capturing, scan denoising, scan alignment, post-processing and mesh reconstruction. Since the capturing process has been discussed in the system overview, this section will focus on the remaining stages.

4.1. Geometric Kinect calibration

The goal of Kinect calibration is: for any given 3D point, its 3D coordinates captured by the two Kinects are aligned in the same reference frame. Different to the traditional calibration methods for multi-view cameras widely adopted in computer vision [12], we make full use of the depth data and resort to the geometry information to perform multi-Kinect calibration. Specifically, the calibration is reduced to a geometry alignment problem and the same process can be reused in the scan alignment stage.

Fig. 3. Denoising example: input (left) and denoised (right).

To perform calibration, we let the two Kinects capture the same object and perform alignment based on the overlapping region of the two partial scans.

To maximize the overlap, we simply place a piece of crumpled-and-then-unfolded A4 paper (shown in the wrapped figure) in-between the two Kinects, serving as the calibrating geometric object. On the one hand, the paper is sufficiently thin, thus the two Kinects are well calibrated when the scans captured at two sides align well. On the other hand, the folded paper contains rich geometric details which is helpful to geometry alignment. Specifically, we employ a supersymmetric third-order graph matching algorithm, detailed in Section 4.3, to perform the rigid alignment of the two scans and calibrate the two Kinects in the geometric way.

4.2. Scan denoising

By adapting the depth filtering technique in [13], we perform consolidation for the partial scans in three steps, i.e., region detection, outlier removal, and spatial filtering. For the region detection step, we employ the Sobel approximation operator to detect the boundary pixels, and remove the unreliable pixels (e.g. pixels near depth edges) by thresholding the depth gradient magnitude, similar to [13]. The main difference between our method and that in [13] is that we only focus on spatial filtering rather than temporal-spatial filtering in [13]. The spatial filtering can denoise the depth data obtained by Kinect effectively as shown in Fig. 3.

4.3. Scan alignment

This step involves pair-wise rigid alignment of the three partial scans obtained by each Kinect. The same technique is also used in the geometry alignment in the former two-Kinects calibration. Perhaps the most well-known pairwise alignment method is the Iterative Closest Points (ICP) method [14]. However, ICP-based algorithms do not work well when the two point clouds to be aligned have small overlap, which is the case of our system since the partial scans captured by one Kinect in different pitching angles do not overlap much. Our alignment method is based on the third-order graph matching technique [15], however, we utilize supersymmetric tensor representing an affinity metric to accelerate the computation. A more generalized formulation for higher order is discussed in [16].

4.3.1. Third-order graph matching with supersymmetric tensor

In our method, point matching is computed by an efficient third-order graph matching approach which takes the advantage of a new compact supersymmetric representation.

Definition 1 *((3-th Supersymmetric Affinity Tensor)).* Given two feature sets P_1 and P_2, with N_1 and N_2 features respectively, let all feature tuples for P_1 and P_2 be F_1 and F_2, there is a matching between feature tuples in F_1 and corresponding tuples in F_2. The supersymmetric affinity tensor is a third-order nonnegative tensor $\mathcal{T}$, for which there exists a set of indices Θ, and a third-order potential function ϕ, such that

$$\mathcal{T}(\Omega(i,j,k)) = \begin{cases} \phi(i,j,k), & \forall (i,j,k) \in \Theta \\ 0, & \forall (i,j,k) \notin \Theta \end{cases} \tag{1}$$

where $i=(i_1,i_2)$, $j=(j_1,j_2)$, $k=(k_1,k_2)$ be pairs of feature points from P_1 and P_2 respectively, $(i_1,j_1,k_1)\in F_1$, $(i_2,j_2,k_2)\in F_2$, Ω denotes an arbitrary permutation of a vector (i,j,k). $\forall (i,j,k)\in\Theta$, $\Omega(i,j,k)\in\Theta$.

A tensor element with $(i,j,k)\in\Theta$ is called a *potential element*, while other elements are called *non-potential elements*. A potential element represents one matching result out of all possible matching candidates. For a set $S_{(i,j,k)}=\{\Omega(i,j,k)\}$, since all the elements have same tensor value based on our new definition, we could represent all potential elements in S by a representative one (i,j,k). We collect all representative elements in a set Γ. Furthermore, as all *non-potential* elements have value zero, there is no need to store them.

Matching between these two feature sets can be represented by an *assignment matrix* $\boldsymbol{X}$ in $\mathcal{A}=\{X\in\{0,1\}^{N_1*N_2}\}$, with each element representing whether a pair $i(i_1,i_2)$ is selected in the matching (if $X_i=1$) or not (if $X_i=0$). From the third-order tensor viewpoint, the matching problem is equivalent to finding the optimal assignment matrix $\boldsymbol{X}^*\in\{0,1\}^{N_1*N_2}$, satisfying [17]

$$\boldsymbol{X}^* = \mathrm{argmax}_{\boldsymbol{X}\in\mathcal{A}} \sum_{i,j,k} \mathcal{T}(i,j,k) X_i X_j X_k. \tag{2}$$

The product $X_iX_jX_k$ will be equal to 1 if the points (i_1,j_1,k_1) are matched to the points (i_2,j_2,k_2), and otherwise 0. $\mathcal{T}(i,j,k)$ is the affinity of the set of assignments $\{i,j,k\}$, which is high if the features in tuple (i_1,j_1,k_1) have similar values to the features in the tuple (i_2,j_2,k_2), and their potential values are similar.

Algorithm 1. Supersymmetric third-order tensor power iteration solution (ℓ^1-norm)

Input: third-order supersymmetric affinity tensor
Output: assignment matrix $\boldsymbol{X}$ with all columns with unit ℓ^1-norm

Initialize $\boldsymbol{X}_0$ to a matrix with all columns with unit ℓ^1-norm, $m=1$
Initialize Y_0 to a matrix with $Y^2_{i,0}=X_{i,0}$
repeat
  set $\widetilde{Y}_m$ to 0
  for each element (i,j,k) of Γ **do**
    $\widetilde{Y}^2_{i,m} = \widetilde{Y}^2_{i,m} + 2\phi(i,j,k)Y^2_{j,m-1}Y^2_{k,m-1}$
    $\widetilde{Y}^2_{j,m} = \widetilde{Y}^2_{j,m} + 2\phi(i,j,k)Y^2_{i,m-1}Y^2_{k,m-1}$
    $\widetilde{Y}^2_{k,m} = \widetilde{Y}^2_{k,m} + 2\phi(i,j,k)Y^2_{i,m-1}Y^2_{j,m-1}$
  end for
  $Y_m(:,c) = \widetilde{Y}_m(:,c)/||\widetilde{Y}_m(:,c)||_2$, $\forall$ column c of Y
  $m=m+1$
until convergence
$X_i = Y^2_i$, $\forall X_i \in X$

The higher-order tensor problem in Eq. (2) can be approximately solved by supersymmetric higher-order power method (S-HOPM), given in [18]. S-HOPM is performed in two iterative steps: higher-order power iteration of $\boldsymbol{X}$, followed by normalization of $\boldsymbol{X}$ under

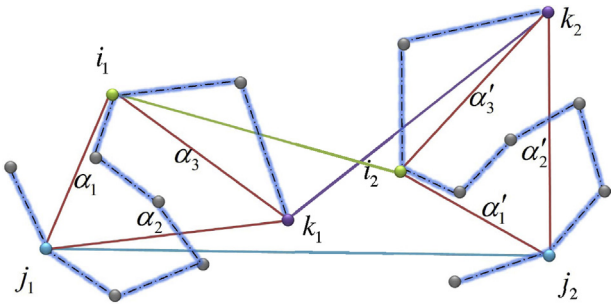


Fig. 4. Illustration of features used in our third-order potential, where edge lengths are invariant to rigid transformation.

the Frobenius norm. As [15], we could relax the set $\mathcal{A}$ to $\mathcal{C}_1$ in which each matrix has unit ℓ^1 norm for all columns. Then we should solve Eq. (3) instead of Eq. (2)

$$\boldsymbol{X}^* = \text{argmax}_{\boldsymbol{X}\in\mathcal{C}_1} \sum_{i,j,k} \mathcal{T}(i,j,k) X_i X_j X_k \tag{3}$$

which is equivalent to solve

$$\boldsymbol{Y}^* = \text{argmax}_{\boldsymbol{Y}\in\boldsymbol{C_2}} \sum_{i,j,k} \mathcal{T}(i,j,k) Y_i^2 Y_j^2 Y_k^2 \tag{4}$$

where $\forall i Y_i^2 = X_i$, $\mathcal{C}_2$ is the matrix set in which each matrix has unit ℓ^2 norm for all columns. Then we could use the S-HOPM algorithm framework for unit ℓ^2 norm to solve Eq. (4). Consider our supersymmetric affinity tensor Definition 1, we would further derive the Y_i expression:

$$\begin{aligned} Y_{i,m}^2 &= \sum_{(i,j,k)\in\Theta} \mathcal{T}(i,j,k) Y_{j,m-1}^2 Y_{k,m-1}^2 \\ &= \sum_{(i,j,k)\in\Gamma} 2\phi(i,j,k) Y_{j,m-1}^2 Y_{k,m-1}^2 \\ &\quad + \sum_{(j,i,k)\in\Gamma} 2\phi(j,i,k) Y_{j,m-1}^2 Y_{k,m-1}^2 \\ &\quad + \sum_{(j,k,i)\in\Gamma} 2\phi(j,k,i) Y_{j,m-1}^2 Y_{k,m-1}^2 . \end{aligned} \tag{5}$$

Eq. (5) are more compact than earlier expressions in the literature, as it handles all symmetrically related potential elements as a single item.

Our third-order potential function ϕ links point feature triples, where triangles formed by three points are *similar* under rotation and translation.

$$\begin{aligned} \phi(i,j,k) &= \phi(\{i_1,i_2\},\{j_1,j_2\},\{k_1,k_2\}) \\ &= \exp(-1/\varepsilon^2 \sum_{(l,l')} \|\alpha_l - \alpha_{l'}\|^2), \end{aligned} \tag{6}$$

where $\varepsilon > 0$ is the kernel bandwidth, which is automatically determined as the average of the ℓ^1 norm of all differences. $\{\alpha_l\}_{l=1}^3$ and $\{\alpha_{l'}\}_{l'=1}^3$ are the edge lengths (in Euclidean distance) formed by feature triples (i_1, j_1, k_1) and (i_2, j_2, k_2): see Fig. 4. The whole solving process is listed in Algorithm 1.

The efficiency of our method benefits from two factors. Firstly, we take the advantage of supersymmetry to derive $\boldsymbol{Y_i}$ as in Eq. (5), using only a single canonical element for computation. Secondly, the power iteration considers only the non-zero potential elements, and excludes each non-potential element from the iteration process. The complexity of the whole iteration process depends only on the number $|\Gamma|$. Consequently, this method also reduces memory cost while keeping accuracy.

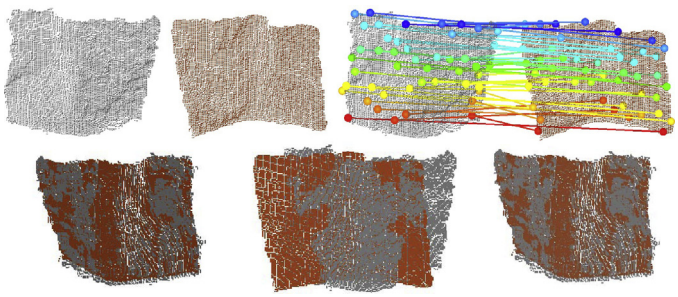


Fig. 5. Alignment of calibration paper. Top row: initial configuration (left) and point matching computed by our method (right); Bottom row: alignment computed by our method (left), by ICP (middle), and by method in [15] (right).

4.3.2. *Pairwise partial scan alignment*

We use the supersymmetric third-order graph matching algorithm to build pairwise matchings based on uniformly sampled feature points. Specifically, a rigid transformation can be computed from each three compatible matching points. Then we use the mean-shift algorithm to cluster the rigid transformations and choose the first model in the clustering space as the optimal alignment transformation.

Fig. 5 demonstrates the two scans capturing both sides of the A4 paper used for calibration, as well as our alignment result. For comparison, the results produced by ICP [14] and the original third-order graph matching algorithm [15] are also shown in the figure. Note that since the two Kinects capture two different sides of the A4 paper, these two scans are mirrored; see Fig. 5 (top-left). For ICP [14], one has to rotate one of the two scans around y-axis by 180° before alignment. Otherwise, ICP [14] would fail to align them as shown in Fig. 5 (bottom-center). However, our algorithm resolves the mirror transformation through computing accurate point matchings. The alignment results produced by our method and the original graph matching method [15] are similar, as reflected by the standard Root-Mean-Square (RMS) errors less than 0.01 m (measured as the nearest point-to-point distances). However, our method takes about 1.0 s to finish the alignment, about half of that taken by the original method [15] (2.3 s). This performance boosting is due to the more compact representation (Eq. (5)) derived from the supersymmetric affinity expression.

Fig. 6 demonstrates the alignment of two partial scans of human body. In this example, our method produces quality alignment while the ICP method [14] clearly fails. This further demonstrates that our method depends less on the initial overlap between the input scans and is more suitable for our task. For this example, our method takes 2.5 s while the original method in [15] takes 6.1 s.

4.4. *Post-processing and reconstruction*

After the alignment, we remove the outlier points of the floor. It mainly remove the floor data. To this goal, we firstly compute the lowest y coordinate, then choose the vertices whose y coordinate

Fig. 6. Alignment of partial scans of human body. From left to right: initial configuration, alignment computed by our method, and by ICP.

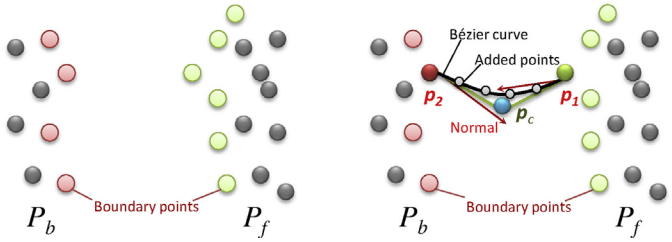


Fig. 7. Illustration of gap filling with Quadratic Bézier curve.

less than $y + 1.5$ cm as the candidate floor points. Secondly, we use the mean shift algorithm for the normal of these candidate points to get the floor points which is the first cluster of the result. Then we estimate a plane P for the floor points. Finally, we treat the points below P and the points whose distance to P is smaller than 1 cm as outliers and remove them.

Since we use two Kinects to capture from the front and back side, there is a narrow band of missing data between the two corresponding partial scans when the in-between regions are invisible from both the two Kinects; see Fig. 8.

We utilize the spatial quadratic Bézier curve to fill the gaps. The basic idea of our gap filling method is illustrated in Fig. 7. We first extract the boundary points of the front and back partial scans, by utilizing the method described in [13], which detected the boundary pixels by thresholding the depth gradient magnitude based on the Sobel operator. Then, we denote them as P_f and P_b, from the front Kinect camera C_f and the back one C_b respectively. The closest point to any point $\mathbf{p}_1 \in P_f$ is searched and denoted as $\mathbf{p}_2$. We build a line whose normal is along the direction from the camera C_f to $\mathbf{p}_1$. Similarly, another line passing through C_b and $\mathbf{p}_2$ is also found. We then compute the closest point to the above two lines, denoted as $\mathbf{p}_c$. Based on the three points $\mathbf{p}_1$, $\mathbf{p}_c$ and $\mathbf{p}_2$, a quadratic Bézier curve is calculated. Finally, we sample the Bézier curve by adding new points whose interval is similar to the average distance of the neighboring points in P_f. Fig. 8 shows the gap filling results for the aligned front and back partial scans, where the newly added points are colored in black, in-between the front (green) and back (red) boundaries.

Based on the captured data and the newly added points in the gap, we utilize the Poisson surface reconstruction algorithm [19] to build the final mesh model. The reconstructed model is a closed triangular mesh with personal features. In Fig. 9, we show that

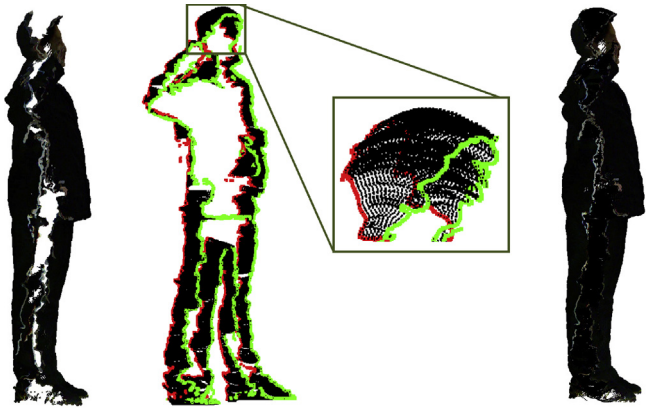


Fig. 8. Gap filling example. Given the front and back scans brought into alignment (left), new points (black) are added to fill the gap between the boundaries of front and back scans, shown in green and red respectively. The final result is shown in the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

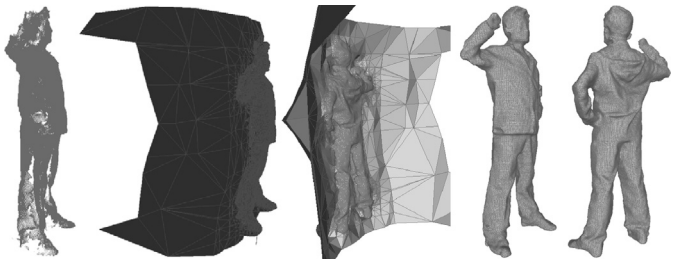


Fig. 9. Poisson reconstruction generates the wrong surface (middle) when applied directly to the aligned partial scans (left). Correct surface can be reconstructed after gap filling (right).

visible artifacts can appear in the reconstruction result if without gap filling.

5. Results

5.1. Experimental results

Our alignment algorithm performs global optimization over all higher-order feature tuples simultaneously to compute point matching, therefore it works well in the small overlap situation, which is difficult to traditional ICP methods. Experiments demonstrate that our method produces desirable results in both the calibration (e.g. Fig. 5) and the alignment (e.g. Fig. 6) stages.

Fig. 9 demonstrates the effect of gap filling on surface reconstruction. The Poisson reconstruction by [19] on the aligned front and back scans presents notable artifacts due to the gap. The final reconstruction result with gap filling is shown in the right.

Fig. 10 demonstrates the capturing and reconstruction of personalized avatar results on different persons. The reconstructed surfaces contain rich geometric details representing personalized features such as face, hairstyle and salient cloth wrinkles. Moreover, since Kinect captures calibrated color images and depth images simultaneously, the final avatar can be automatically textured with the color images as shown in the right two columns in Fig. 10.

5.2. Evaluation

In summary, our system possesses the following features which are critical for practical usability.

- Flexible and simple system setup. In our system, the configuration of the two Kinects is quite flexible and easy to do for a casual user.
- Low price. A Microsoft Kinect is at a price of about $150, so the total cost of our system is about $300.
- Small space occupation. The space occupation of our system is about 3 m × 1 m × 2 m which is much smaller than that of [5].
- Fast capture. The capturing process is fast so that the user hold a static pose about 3 s.

The running time (tested on a PC laptop with Intel Core i7 processor at 1.6 GHz) is reported in Table 1. The whole process of capturing and reconstruction takes about 1 min.

Six biometric measurements are calculated on the constructed human models and compared with those measured on the corresponding real persons. The accuracy of our biometric measurements is comparable to that of the methods in [5] and [4].

Table 1
Average running time (s).

Capture	Calibration	Denoising	Alignment	Reconstruction	Texture
3	1	2	10	40	10

Fig. 10. Capturing and reconstruction of personalized avatar for various persons. From left to right: aligned front scans, aligned back scans, merged front and back scans, reconstructed surface, textured model in two views (grey regions depict filled gaps).

Table 2
Average error of biometric measurements (cm).

Height	Neck to hip distance	Shoulder width	Arm length	Leg length	Waist girth	Hip girth
1.0	2.4	1.9	3.2	2.2	6.5	4.0

Table 3
Performance comparison with existing methods.

	[4]	[5]	[8]	Ours
Devices	1 Kinect	3 Kinects + 1 turntable	1 Kinect	2 Kinects
T_c	Over 4 s	30 s	30 s	3 s
T_p	65 min	6 min	5 min	1 min
Accuracy	2.2 cm	6.2 cm	4 cm	6.5 cm

Table 2 shows the average error in centimeter. It can be seen that the reconstructed models approximate the real persons well.

The accuracy of the reconstructed mesh is mainly affected by two aspects: geometry capture and hole filling. We would explain them in the following:

- The geometric capture accuracy is determined by the Kinect device, whose performance has been thoroughly analyzed by [20]. In a short, the accuracy regresses quadratically with the distance measurement. Kinect's depth accuracy is from accurate to 2 mm at 1 m (3.3 ft) distance from Kinect to accurate to 2.5 cm at 3 m (9.9 ft) distance from Kinect. The distance between two Kinects is 1.5–3m, this implies that the distance between the capturing subject and one Kinect is about 0.75–1.5m, so the depth error of our system would be less than 5 mm for the captured data.
- The hole filling, a key step to reconstruct the 3D mesh, is achieved by the spatial quadratic Bézier curve fitting. The new added points are generated from the simulated curves, rather than captured in the real way. This would definitely add reconstruction error, which is larger than those resulting from the capture. It is also easy to understand the average error difference of biometric measurements. The waist part in the large volume always has loose and fat belly, so, the waist girth has the largest error due to the hole filling from the boundary points.

5.3. Comparison

A clean comparison, between our system and existing methods [4], [5], and [8], is shown in Table 3, where T_c and T_p are the capture and processing time respectively, The accuracy is the maximal L_2 distance error. It is noticeable that the key feature of our system is the capture speed, a moment, or quantitatively 3 s, which is short enough for the person being captured to hold a static pose stably and comfortably.

5.4. Skeleton-driven avatar animation

The reconstructed personalized avatar can be animated by using the precomputed skeleton provided by Kinect. Alternatively, we can extract skeleton using the method of [21]. As a preprocessing, we utilize the method in [22] to automatically rig the extracted skeleton to the reconstructed avatar. Fig. 11 demonstrates the animated avatar driven by the Kinect skeleton.

Fig. 11. Skeleton-driven avatar animation, with the rigged skeleton shown in green. Since Kinect mirrors the real scene, the output avatar is mirrored w.r.t. the person being captured. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Fig. 12. Two failure cases where the bodies are occluded (red circle) and the reconstructed geometry is incorrect. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

5.5. Failure case

Our gap filling only works for the missing parts between the front and back scans, but not for general missing data due to occlusion. When the user takes a static pose with self-occlusion with respect to both Kinects, our method does not provide a mechanism to recover the occluded geometry. Fig. 12 shows two failure cases where the bodies are occluded and the reconstructed geometry is incorrect.

6. Conclusions

We present a system for fast capture of personalized avatar of any person, which has been tested and shown to be effective in out-of-laboratory environment. The experimental results demonstrate that our system is easy-to-use with a relatively simple system setup (two easily configured Kinects), low price, small space occupation, and fast capturing speed (in 3 s). The capture and reconstruction process is fully automatic, taking about 1 min to obtain the full result.

There are several aspects for further development to improve our system. Firstly, the quality of the reconstructed models is still not high due to the low quality depth data obtained by Kinects. More advanced methods for point cloud denoising and super-resolution could be utilized to better consolidate the depth data. Secondly, one of the main challenges of point cloud precessing is how to fill complex holes in noisy depth data with plausible surfaces. It is more natural and effective to complete the missing regions with the help of prior template. Finally, more studies could be conducted to further improve the usability of the capturing system by exploiting the new features in the latest release of Kinect SDK.

Acknowledgments

This work was supported by the Natural Science Foundation of China (Nos. 61103084, 61272334, 61202333) and CPSF (2012M520392).

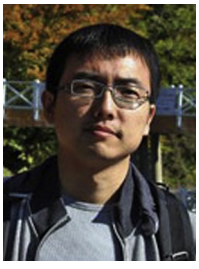
References

[1] Kinect. Microsoft; 2012 http://www.xbox.com/kinect
[2] Anguelov D, Srinivasan P, Koller D, Thrun S, Rodgers J, Davis J. Scape: shape completion and animation of people. ACM Transactions on Graphics (Special Issue of SIGGRAPH) 2005;24(3):408–16.
[3] Cyberware. The cyberware whole body color 3d scanner; 2012 http://www.cyberware.com/products/scanners/wbx.html
[4] Weiss A, Hirshberg D, Black MJ. Home 3d body scans from noisy image and range data. In: International conference on computer vision. 2011. p. 1951–8.
[5] Tong J, Zhou J, Liu L, Pan Z, Yan H. Scanning 3d full human bodies using kinects. IEEE Transactions on Visualization and Computer Graphics 2012;18(4):643–50.
[6] Stern DN. The Present Moment: In Psychotherapy and Everyday Life. New York: W W Norton Co.; 2004.
[7] Izadi S, Kim D, Hilliges O, Molyneaux D, Newcombe R, Kohli P, Shotton J, Hodges S, Freeman D, Davison A, Fitzgibbon A. 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. 2011. p. 559–68.
[8] Cui Y, Stricker D. 3d shape scanning with a kinect. In: ACM SIGGRAPH 2011 Posters. 2011.
[9] Nagy E. Sharing the moment: the duration of embraces in humans. Journal of Ethology 2011;29:383–91.
[10] Chen Y, Cheng Z-Q. Personalized avatar capture using two kinects in a moment. In: SIGGRAPH ASIA Posters 2012. 2012.
[11] OpenNI. Openni organization; 2012 http://openni.org/
[12] Svoboda T, Martinec D, Pajdla T. A convenient multi-camera self-calibration for virtual environments. PRESENCE: Teleoperators and Virtual Environments 2005;14(4):407–22.
[13] Richardt C, Stoll C, Dodgson NA, Seidel H-P, Theobalt C. Coherent spatiotemporal filtering, upsampling and rendering of RGBZ videos. Computer Graphics Forum 2012;31(2):247–56.
[14] Rusinkiewicz S, Levoy M. Efficient variants of the ICP algorithm. In: International conference on 3D digital imaging and modeling. 2001. p. 145–52.
[15] Duchenne O, Bach F, Kweon I-S, Ponce J. A tensor-based algorithm for high-order graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence 2011;33:2383–95.
[16] Cheng Z-Q, Chen Y, Martin RR, Lai Y-K, Wang A-P. Supermatching: feature matching using supersymmetric geometric constraints. IEEE Transactions on Visualization and Computer Graphics 2013;19(11):1885–94.
[17] Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Review 2009;51(3):455–500.
[18] Kofidis E, Regalia PA. On the best rank-1 approximation of higher-order supersymmetric tensors. SIAM Journal on Matrix Analysis and Applications 2002;23(3):863–84.
[19] Kazhdan M, Bolitho M, Hoppe H. Poisson surface reconstruction. In: Eurographics symposium on geometry processing. 2006. p. 61–70.
[20] Khoshelham K, Elberink SO. Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 2012;12(2):1437–54.
[21] Shotton J, Fitzgibbon AW, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. Real-time human pose recognition in parts from single depth images. In: CVPR. 2011. p. 1297–304.
[22] Baran I, Popović J. Automatic rigging and animation of 3d characters. ACM Transactions on Graphics 2007;26(3).

Yin Chen received a BSc and MSc degree from School of Computer at National University of Defense Technology in 2008 and 2010, respectively. He is a PhD student, School of Computer at National University of Defense Technology. His research interests include computer graphics, and digital geometry processing.

Gang Dang received a BSc, MSc and PhD degree from School of Computer at National University of Defense Technology in 1995, 1997 and 2001, respectively. He is an associate professor, School of Computer at National University of Defense Technology. His research interests include computer graphics, and virtual reality.

Zhi-Quan Cheng received a BSc, MSc, and PhD degree from School of Computer at National University of Defense Technology in 2000, 2002 and 2008, respectively. He is a researcher at Avatar Science Company. His research interests include computer graphics, and visual computing.



Kai Xu received a BSc, MSc and PhD degree from School of Computer at National University of Defense Technology in 2004, 2006 and 2011, respectively. He is a lecture, School of Computer at National University of Defense Technology. His research interests include computer graphics, and digital geometry processing.