Signal Processing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Signal Processing

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

# 3D real human reconstruction via multiple low-cost depth cameras

Zhenbao Liu, Hongliang Qin, Shuhui Bu *, Meng Yan, Jinxin Huang, Xiaojun Tang, Junwei Han

*Northwestern Polytechnical University, China*

ARTICLE INFO

*Article history:*
Received 12 March 2014
Received in revised form
12 October 2014
Accepted 14 October 2014

*Keywords:*
3D real human
Reconstruction
Low-cost depth camera

ABSTRACT

In traditional human-centered games and virtual reality applications, a skeleton is commonly tracked using consumer-level cameras or professional motion capture devices to animate an avatar. In this paper, we propose a novel application that automatically reconstructs a real 3D moving human captured by multiple RGB-D cameras in the form of a polygonal mesh, and which may help users to actually enter a virtual world or even a collaborative immersive environment. Compared with 3D point clouds, a 3D polygonal mesh is commonly adopted to represent objects or characters in games and virtual reality applications. A vivid 3D human mesh can hugely promote the feeling of immersion when interacting with a computer. The proposed method includes three key steps for realizing dynamic 3D human reconstruction from RGB images and noisy depth data captured from a distance. First, we remove the static background to obtain a 3D partial view of the human from the depth data with the help of calibration parameters, and register two neighboring partial views. The whole 3D human is globally registered using all the partial views to obtain a relatively clean 3D human point cloud. A complete 3D mesh model is constructed from the point cloud using Delaunay triangulation and Poisson surface reconstruction. Finally, a series of experiments demonstrates the reconstruction quality of the 3D human meshes. Dynamic meshes with different poses are placed in a virtual environment, which can be used to provide personalized avatars for everyday users, and enhance the interactive experience in games and virtual reality environments.



## 1. Introduction

Real 3D scene and human reconstruction from a professional 3D scanner has been studied extensively in multimedia, virtual reality, and computer graphics [1]. Satisfactory results have been obtained that are adaptable to industrial inverse engineering and production designs based on virtual reality. However, the techniques cannot be applied directly in home-centered environments owing to the high cost, large volume, complex operation, and computational burden. In recent years, portable low-cost, easy-to-use, RGB-D cameras such as the Kinect [2] have become very popular and widely used. However, this type of camera captures low-quality depth images, which is a major constraint in providing high-quality immersive and virtual applications. Thus, many researchers have shown great interest in recent developments in scene reconstruction and human modeling using RGB-D sensors.

There are several pioneering works focusing on interesting 3D virtual applications of RGB-D sensors. Alexiadis et al. [3] built a real-time automatic system for dance performance

* Corresponding author. Tel./fax: 86 29 88492344.
*E-mail addresses:* liuzhenbao@nwpu.edu.cn (Z. Liu), bushuhui@nwpu.edu.cn (S. Bu).

http://dx.doi.org/10.1016/j.sigpro.2014.10.021
0165-1684/

evaluation using a Kinect RGB-D sensor, and provided visual feedback for beginners in a 3D virtual scene. Liu et al. [4] used depth sensor data from a Kinect to track human movement and evaluate players' energy consumption to investigate how much energy is expended while playing in a virtual environment. Bleiweiss et al. [5] proposed a solution to animate in-game avatars using real-time motion capture data, and blend actual movements of players using predefined animation sequences. Pedersoli et al. [6] provided a framework for a Kinect that enables more natural and intuitive hand-gesture communication between a human and a computer. Tong et al. [7] presented a novel scanning system for capturing different parts of the human body at close range, and then reconstructing a whole 3D human body. However, this system requires external calibration and an accurate capture space, which is not available in home applications. The system proposed by Anguelov et al. [8] deforms a data-driven shape completion and animation of people (SCAPE) model to fit the scanned data given a limited set of markers specifying the target shape. Using a single static scan and markers for motion capture, their system constructs a high-quality animated surface model of a moving person with realistic muscle deformation. Similarly, Weiss et al. [9] introduced a SCAPE model with 15 body parts that fits poses of a real human to a 3D model, while Chen et al. [10] adopted training data, including 3D meshes for multiple users with different poses as obtained from SCAPE, to model existing 3D human body meshes. Alexiadis et al. [11] proposed a full 3D reconstruction of moving foreground objects from depth cameras. Several authors have also focused on other related applications, e.g., tracking a human body [12,13], recognizing human poses in real time [14], and converting a movie clip into a comic [15] using depth cameras.

However, unlike commercial 3D scanners that can output clean, dense point clouds for use in building a structured mesh easily, the point cloud captured from depth cameras is sparse, discontinuous, and noisy, resulting in the generation of many holes and missing depth values. This is caused by the nature of infrared light measurement, which is affected by multiple factors, including scattering, absorption by dark objects, transmission through transparent objects, multiple reflections, and large distances. We propose a pipeline aimed at obtaining a real textured 3D human body from low-cost depth cameras. We first remove the background to obtain a 3D partial view of the human from each depth image with the help of intrinsic and stereo calibration parameters, and successively register two neighboring partial views as a concatenated partial view. The complete 3D point cloud of the human is globally registered using all the pairwise registrations. A 3D mesh structure incorporating color is reconstructed using Delaunay triangulation and the Poisson method. To verify the effectiveness, efficiency, and robustness of the real 3D human reconstruction using multiple low-cost depth cameras, we built an experimental environment. Six depth cameras (Kinect) were placed at different positions on a circle to capture multiple views with as many overlapping points as possible. Fig. 1 shows the entire process from setting up the hardware system to the final 3D human reconstruction. The results show that the quality of the reconstructed mesh is satisfactory.

How to build a correspondence of partial 3D views automatically and effectively is the key to 3D human reconstruction. Shape correspondence, matching, and retrieval of 3D shapes have been extensively investigated in recent years [16–18], and these are still hot topics in multimedia, computer vision, computer graphics, and computer-aided design. Many researchers have attempted to provide content-based matching and retrieval techniques, focusing mainly on local point description and high-level feature extraction [19–22], topological structure [23], non-rigid shape feature [24], sketch based retrieval [25], view comparison [26–31], and a relevance feedback mechanism [32]. For example, to match a single-view Kinect scan to high-quality 3D models, Shen et al. [33] proposed recovering the underlying structure of a scanned object by assembling suitable parts obtained from the repository models. Kim et al. [34] presented an efficient method for acquiring 3D indoor environments with variability and repetition through a Kinect sensor. Their work segments a single 3D point cloud scanned using a real-time simultaneous localization and mapping (SLAM) technique, classifies it into plausible objects, recognizes these using primitive fitting and connected component analysis, and extracts their pose parameters. Kakadiaris et al. [35] used a composed deformable model to fit 2D body silhouettes extracted from three mutually orthogonal views. Wang [36] constructed a feature wireframe of a human body from laser scanned 3D unorganized points using semantic feature extraction, and modeled the symmetric detail mesh surface of the human body. Hsieh et al. [37] presented a novel segmentation algorithm to segment a body posture into different body parts using the deformable triangulation technique. Holte et al. [38] summarized recent approaches for 3D human pose estimation based on 3D features reconstructed from multiple views. However, our problem is different: we require robust correspondence between noisy partial views without markers, which are scanned by low-cost depth cameras. Moreover, global registration with a small accumulated error distributed among multiple cameras must be automatically implemented in almost real time. We design a registration framework to realize the two objectives simultaneously.

*Contributions*: The past decade has seen the emergence of new 3D imaging devices and techniques capable of capturing full human body shapes [39], which are used extensively used in human engineering, health assessment, protective equipment (car or airplane seats), medical diagnosis, entertainment, the clothing industry, and virtual reality. In this study, we implemented an entire reconstruction pipeline for a real 3D human as a prototype system composed of multiple low-cost depth cameras. Furthermore, the effectiveness, efficiency, and robustness of the system are investigated by applying it to a group of users with different poses. The entire process is completely automatic and does not require human body landmarks, or any manual operation on the 3D model, which is important for dynamic virtual reality environments. Our approach is realizable with the following two technical contributions:

1. A solution to 3D partial view generation of humans is proposed. Given color images and noisy depth images captured by a single sensor, our method adopts background

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

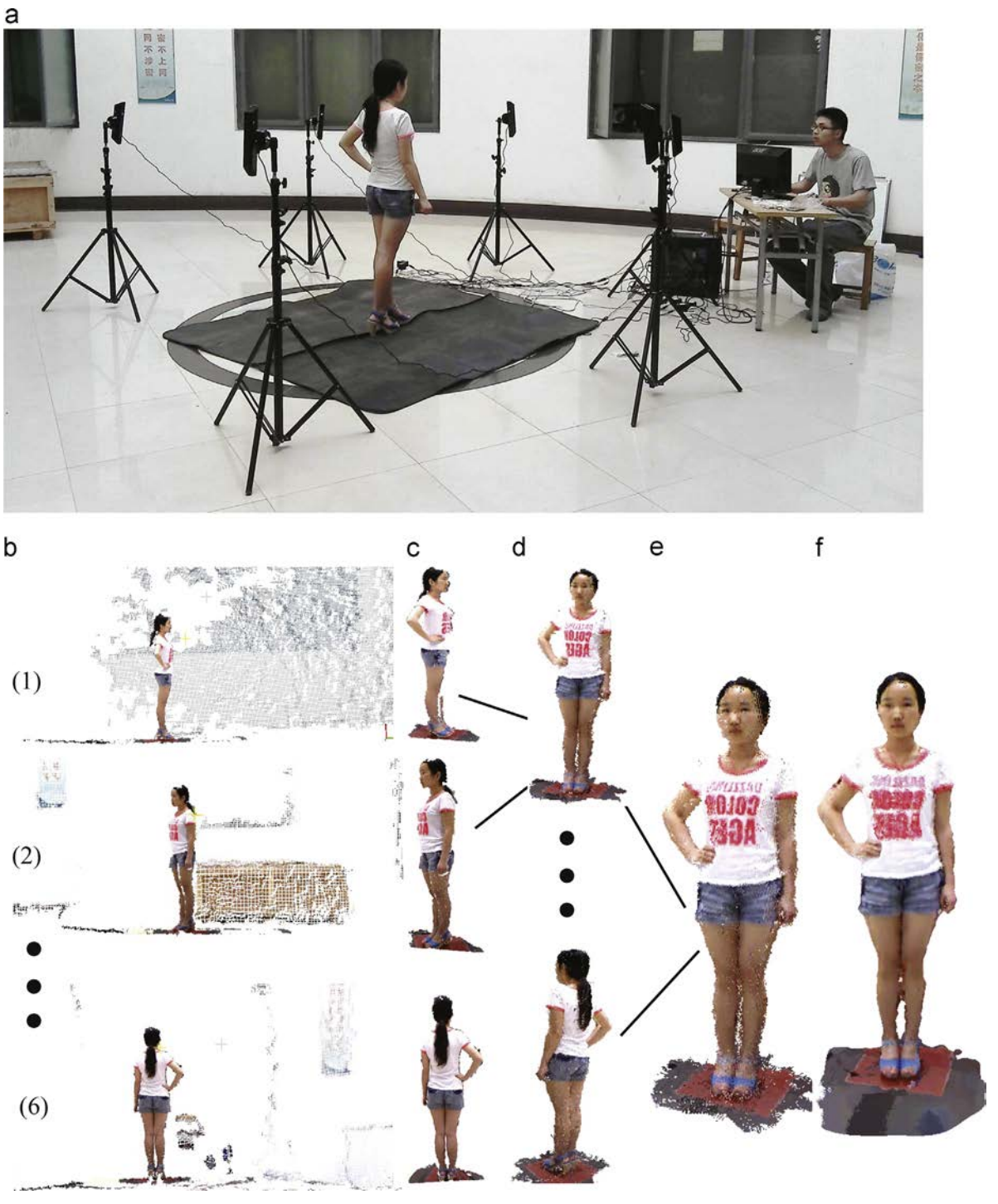


**Fig. 1.** Overview of the steps in 3D human reconstruction for our original data. (a) Our experimental environment where six depth cameras are placed. (b) A single scene shot via one depth camera. (c) The obtained 3D partial view after background removal. (d) New 3D partial view after pairwise registration. (e) The globally registered 3D human. (f) The reconstructed 3D mesh model.

removal to extract a coarse depth view of the human, followed by depth smoothing. Partial view filters are employed to obtain a relatively clean 3D partial point cloud.

2. To provide more robust and accurate registration for sparse and noisy point clouds, we introduce an initial pairwise registration method using segment constrained correspondence. The correspondence is built for key points and uniform sampling points, and performed by comparing their feature descriptors. We add a segment constraint to spectral matching particularly for a human point cloud, which significantly

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

improves the correspondence effect. The proposed scheme provides an initial alignment around the optimal solution of fine registration.

## 2. 3D partial view generation

In this section, we propose a method for extracting 3D partial views of humans from captured RGB-D images to obtain a relatively clean 3D human automatically from cluttered environments. The scheme relies on dealing with noisy low-resolution depth maps by resorting to relatively high-quality color images. The static background in a color image is easily removed and mapped to the depth image to extract a depth view of the human. Internal calibration parameters are used to convert the depth view to a coarse 3D point cloud of the human. Further, we employ partial view filters to smooth the point cloud as much as possible.

*Background removal*: We assume that the background is static when the target for capture enters the environment. Our method is able to detect and track the person's movement, making it feasible to use background subtraction. During the tracking process, a pixel in the color image is labeled as "in motion" if its color differs significantly from pixels in the background. This classification is taken from adaptive background mixture models [40], where each pixel is represented as a mixture of Gaussians. A background model is trained on several initial static background images. The estimation problem of model parameters is solved by the expectation maximization method, while the number of Gaussian mixture models is adaptively set to enhance robustness against noises. After the model has been constructed, we infer for each pixel in the dynamic environment whether the pixel belongs to the person or the background. Since the background is required a prior, it is easy for depth cameras to capture a few background images in advance. In our implementation, we only use 10 images to construct the background model; thus, less than a second is needed to capture the background. Both background modeling and human extraction are completed in real time. After obtaining a 2D human from the color images, the depth maps are combined with a human mask using a rotation and translation matrix determined by stereo calibration. We then obtain the 3D point cloud of the captured human. Fig. 2 shows the effect of background removal.

*Depth smooth*: Accuracy of depth images is a big problem in low-cost depth cameras that produce images with noise, low resolution, and especially holes. These holes originate mainly from the basic measurement principle. Infrared light measurement is influenced by scattering, transparent and dark objects, and multiple reflections. If the missing depth values are not recovered, a low-quality 3D reconstruction is obtained. Since these missing depth values tend to appear in the form of groups, we employ a recursive median filter [41] to fill the holes. Compared with the traditional median filter that considers all neighboring pixel values, we take the median of the non-missing values in a local grid centered on the current pixel. This median filter is recursively applied until all the missing values have been filled. In practice, only a few iterations are needed, and all the values are recovered quickly.

*Partial view filter*: We find that the obtained 3D partial views are affected by noise and outliers. Our filter includes two steps: a coarse band pass filter and a fine Gaussian filter. First, we use a band pass filter $F(b)$ to obtain a coarse point cloud for the partial view. This is expressed as follows:

$$F(b)*p_i = \begin{cases} p_i, & x \in [x_0, x_1], y \in [y_0, y_1], z \in [z_0, z_1] \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\{[x_0, x_1], [y_0, y_1], [z_0, z_1]\}$ denotes the bounding box of the human point cloud, $p_i$ is a point in the point cloud, and $F(b)*p_i$ denotes applying the filter to point $p_i$. Here, the bounding box is constructed with a recent high-level skeleton tracking algorithm [14]. The algorithm designs an intermediate body part representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. The classifier is trained by efficient randomized decision forests, considering a large number of characteristics, such as short or tall and thin or fat, and many poses, such as models with different hair styles and clothing. Once we obtain the positions of the tracked

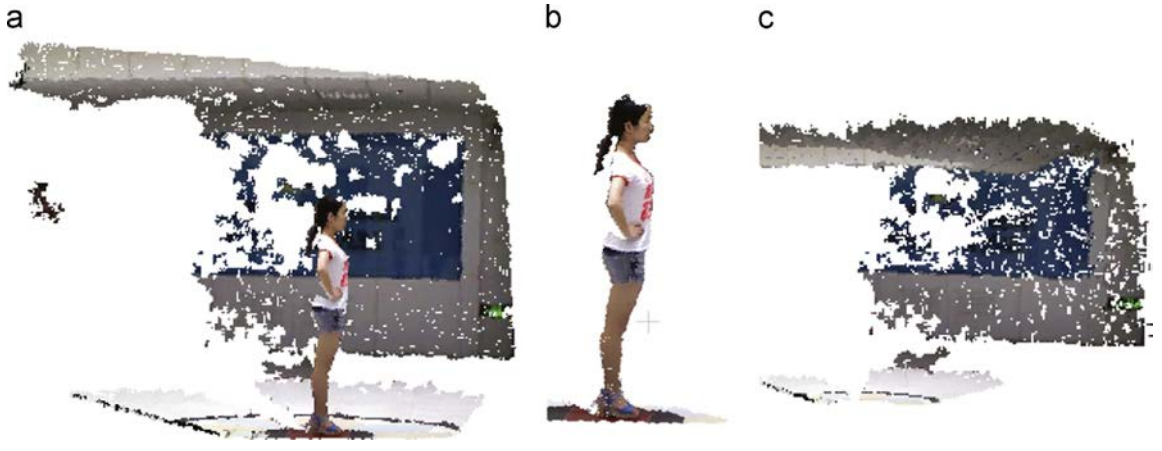


**Fig. 2.** (a) 3D scene scanned by a single depth camera. (b) The human extracted from cluttered background. (c) The removed background.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

skeleton joints, the bounding box of the whole skeleton is simply computed by utilizing the maximum and minimum values of all the joint coordinates. We empirically enlarge the bounding box of the skeleton by 10%, and set the bounding box of the human as the band pass filter $F(b)$.

Assuming that the noisy data obey the Gaussian distribution, we use an iterative Gaussian filter to obtain a clear boundary. First, we search the nearest neighbors of each 3D point, and compute their mean and standard deviation. Points beyond two standard deviations are considered as noise. The procedure is repeated several times to obtain a relatively smooth 3D point cloud. This process is illustrated in Fig. 3.

## 3. Feature correspondence for initial registration

Latest advances dealing with feature correspondence have mainly focused on the following topics: feature extraction based on a local convolutional auto-encoder [42] and topic model [43], and feature comparison via Hausdorff distance learning [44]. Differing from these studies, our work relies on key point selection on a 3D point cloud, and computes feature descriptions for coarse correspondence between partial 3D views.

*Key point selection*: Owing to the large number of points in a point cloud, we perform the computations on subsets. The subsampling is carried out using a combination of key point sampling and uniform sampling. A key point detector describes the distinctive information in its local region. It is desirable that regions with visual saliency such as protrusion, tip, and transition between different parts have larger descriptive values, while flat or smooth regions have smaller values. Coding these regions according to saliency is the task of the key point detectors. In this regard, there has been some early research like splash features, shape indexes, and point signatures. Here we investigate the recently proposed 3D SIFT key point detector [45] extended from 2D SIFT [46] where the detected features are highly distinctive and invariant to image scale and rotation. First, we compute the local maximum of 3D Hessian determinant, and then a number of Gaussian filters with different deviations are convolved with the input data. Key points are identified as local minima of the difference of Gaussians. To consider the necessary spatial uniformity we also perform positional subsampling by constructing an octree and picking one sample point in each voxel. Fig. 4 shows the results of key point detection and uniform sampling.

*Feature descriptor*: Describing the feature of a local 3D structure and differentiating it from other points mainly depend on a descriptor with discriminative power [47]. While state-of-the-art approaches in computer vision build a visual vocabulary [48,49] based solely on visual statistics of local image patches, they cannot be applied to solve our problem owing to the lack of a large training dataset for human reconstruction. Moreover, it is particularly important to find a 3D feature descriptor that is robust to incomplete data, clutter, and sensor noise. After extensive investigation, we chose the recent Signature of Histograms of Orientations (SHOT) [50] as our feature descriptor. The SHOT descriptor is based on a repeatable local reference frame, using eigenvalue decomposition around a given input point. Then a spherical grid centered on the point divides the neighborhood such that in each grid bin a weighted histogram of normal is obtained. The descriptor is computed by concatenating all histograms into the final signature. The length of the signature is 352 since 11 shape bins and 32 divisions of the spherical grid are adopted in the reference frame encoded by nine parameters. The descriptor is scaled to make its norm equal to one. Because the acquired 3D partial views contain the same texture information, consideration of texture information boosts the descriptive power. The inclusion of texture cues [51] results in more effective descriptions, without negatively affecting on efficiency. The descriptor with color information has 1344 dimensions. Fig. 5 shows two types of feature descriptors for a given red point on a 3D partial view.

*Segment constrained correspondence*: Graph-based matching and learning have attracted great research interest in recent years. For example, Wang et al. [52] investigated a multi-graph learning method for video annotation. Several graph-based learning approaches have been applied for web image search [53] and image indexing [54]. In this work, the proposed feature correspondence of two partial 3D views follows the spirit of the influential work by Leordeanu et al. [55], who employed the spectral properties of the weighted adjacency matrix $M$ of a graph. Each potential point pair is regarded as a node in the graph. Both the inter-similarity and the inner geometric constraint of point pairs are considered for correspondence. In our segment constrained correspondence, which differs from the method described above, only a pair of points belonging to corresponding segments in two views qualifies as a candidate pair. This greatly reduces the number of candidate pairs. Moreover, evidently incorrect pairs, but with similar descriptions are filtered out. Therefore, partial 3D data are pre-segmented before correspondence, and the segmented part is seen as prior information in the correspondence stage. Because a human in each acquired 3D partial view always appears straightforwardly upright, spatially segmented parts are added through height clustering as strong constraints for spectral matching. Specifically, each point correspondence pair $p$ should belong to the same cluster $C$, and therefore, we incorporate a cluster indicator function $c(x)$ into the adjacency matrix

$$c(p)=\begin{cases}1, & (x,y)\in C\\ infinity, & (x,y)\notin C\end{cases} \tag{2}$$

$$m(p,q)=\begin{cases}\exp\left\{-\dfrac{d^2(p)}{2\sigma^2}\right\}+\lambda c(p), & p(x,y)=q(x',y')\\ \exp\left\{-\dfrac{(d(x,x')-d(y,y'))^2}{2\sigma^2}\right\}, & p(x,y)\neq q(x',y')\end{cases} \tag{3}$$

where $d(p)$ denotes the feature distance between the point pairs in different views, $d(x,x')$ denotes the spatial Euclidean distance between two points in the same 3D view, and $\lambda$ is a tradeoff parameter determined experimentally.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021


**Fig. 3.** (a) 3D view without iterative Gaussian filter. (b) 3D view after iterative Gaussian filter. (c) Removed noisy points.

The desired correspondence is selected from a large number of candidate pairs. Given the initial correspondence solution $P$, the optimal solution $P^*$ is obtained by optimizing the following quadratic polynomial problem:

$$P^* = \arg\max P^T M P \quad \text{s.t.} \quad P \in \{0,1\}^n. \tag{4}$$

The problem can be effectively solved by quadratic pseudo-boolean optimization (QPBO) [56]. The solution is a set of indicator values representing the selected correspondence pairs. Fig. 6 shows a comparison of our method and spectral matching [55]. Using the method given above, initial correspondences are automatically

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

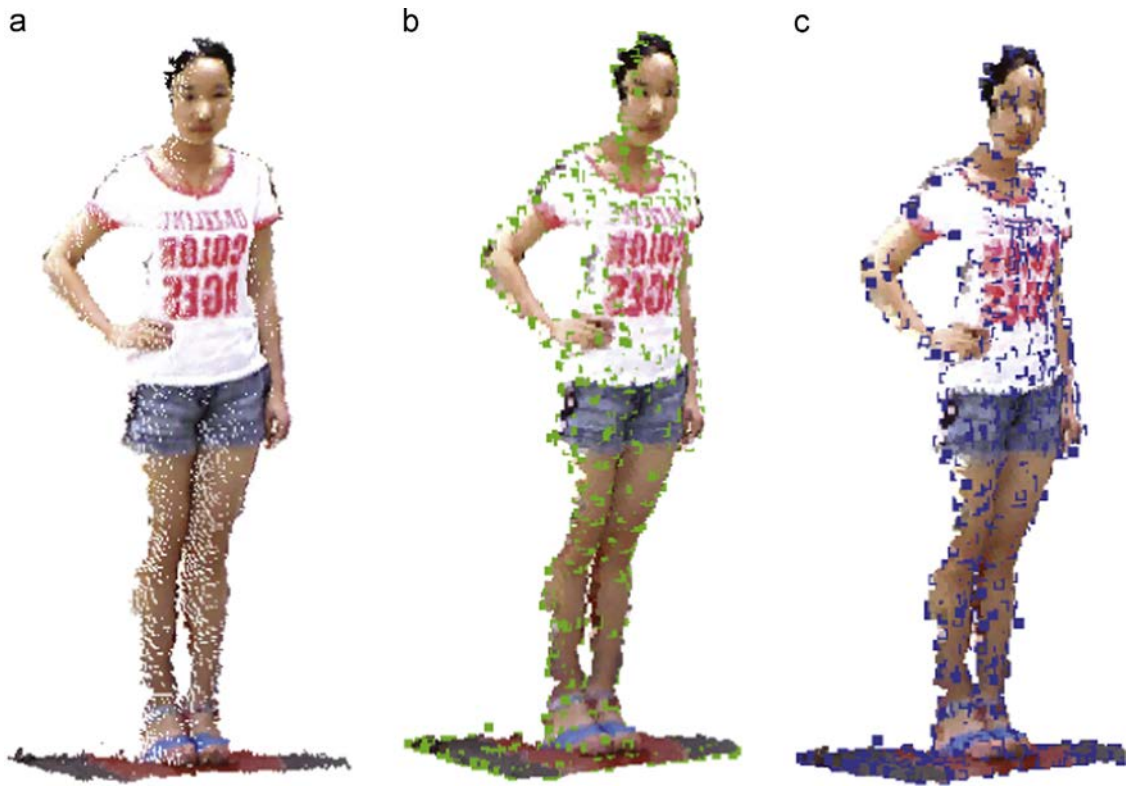


**Fig. 4.** (a) 3D partial view. (b) Uniform sampling based on octree. (c) 3D SIFT key points.

generated. Since this results in a coarse rigid transformation only, it can be accurately refined by considering the spatial Euclidean distance between each pair in the fine pairwise registration.

## 4. Fine pairwise registration

To obtain accurate registration of two partial 3D views and significantly reduce the influence of noisy data and outliers, we follow the notion of coherent point drift [57]. A set of points $X$ in one partial 3D view can be modeled using Gaussian mixture models, while another set of points $Y$ is regarded as the data generated by these Gaussian mixture models. The posterior probability of data points in $Y$ corresponding to the Gaussian mixture models centered on points in $X$ should be maximized after transforming the source points $X$ to align with the target points $Y$. The transformation is represented by rotation $R$ and translation $t$ in this case. Accurate correspondences between two sets of points and rigid transformations are found by optimizing the maximum likelihood estimation

$$E(R,t)=\frac{1}{2\sigma^2}\sum_X\sum_Y P(x_i|y_j)\|y_j-Rx_i-t\|^2+\frac{N_pD}{2}\log\sigma^2 \qquad (5)$$

where $P(x_i|y_j)$ is the posterior probability that the set of points $Y$ corresponds to the set of points $X$, and $N_p$ is the sum of the posterior probability of all point pairs. $D$ is the dimension of the data, which is set to three. Similar to the expectation maximization method, we first fix the transformation and compute the posterior probability using Bayes' theorem, and then update the transformation by maximizing the likelihood function mentioned above. These two steps are iteratively executed until convergence. Given two partial 3D views captured by depth cameras, despite containing a large amount of noise and several holes, an accurate registration result can be achieved using our method, as illustrated in Fig. 7. Since multiple neighboring views contain overlapping information, we sequentially perform pairwise registration for any two neighboring views.

## 5. Global registration

While sequentially aligning each 3D partial view pairwise with its neighbor, the error gradually accumulates so that the starting and ending partial views cannot be aligned. We introduce global registration to obtain a consensus registration for all partial views by diffusing pairwise registration errors. From all the pairwise registrations, we iteratively select the view with the most matching pairs. The current view is aligned with all its neighboring views to obtain the new registration error. If the error exceeds a threshold, another view is selected from the neighbors to reduce the error once again. The new chosen view is successively aligned with its neighboring views to compute the current registration error. This process ensures that larger pairwise registration errors are reduced, and smaller errors increase simultaneously. Finally, these pairwise registration errors are distributed evenly among all views.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

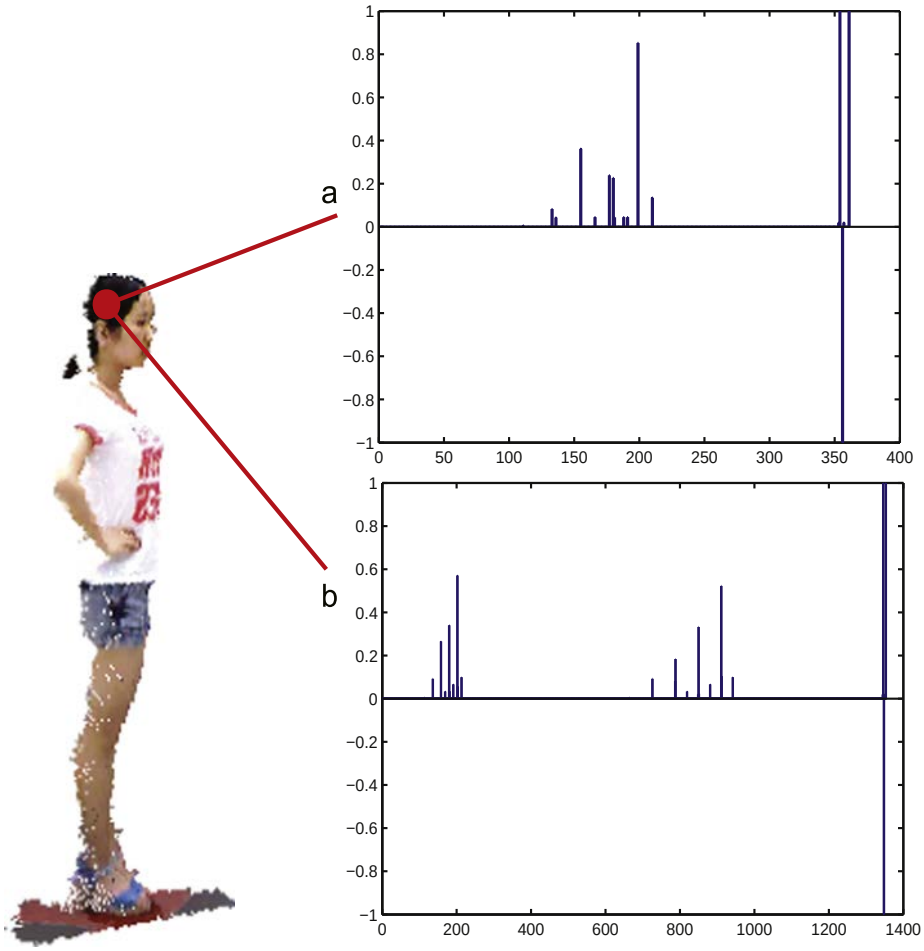


**Fig. 5.** The feature descriptors for a given red point on 3D partial view. (a) SHOT feature. (b) SHOT feature with color information. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

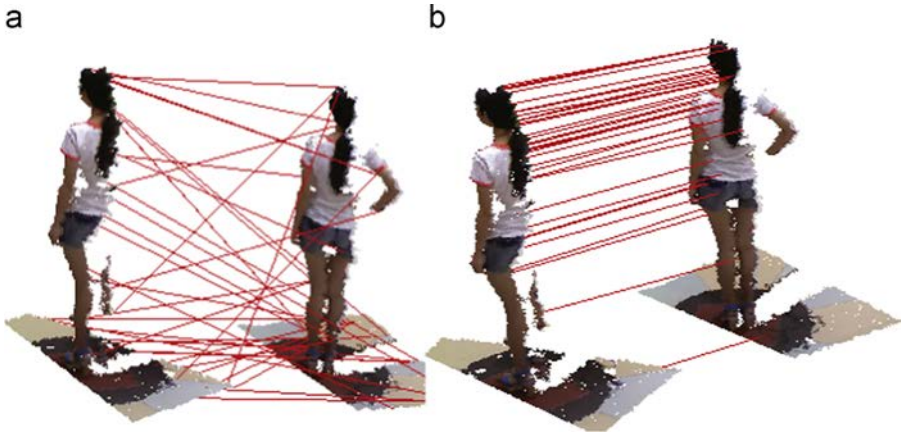


**Fig. 6.** (a) Correspondence based on spectral matching. (b) Segment constrained correspondence. It can be seen that there are many correspondence errors in (a) and these errors are corrected in our method (b).

## 6. Human mesh reconstruction

The human point cloud including color generated above has three defects, which may cause an incorrect mesh reconstruction. One problem is that the point cloud is not smooth and contains several burrs on the boundary, resulting from interference of several RGB-D cameras, inevitable pairwise and global registration errors, and the background mixture. We solve the organization problem of these unordered 3D points, and adopt a polygonal form

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

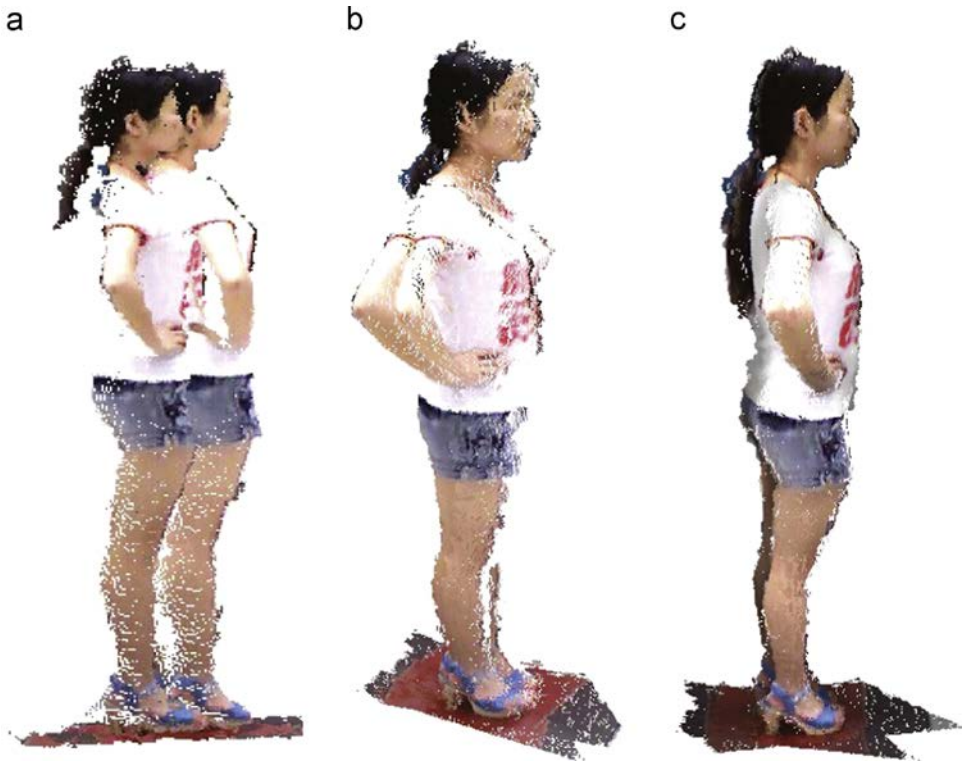


**Fig. 7.** (a) Two partial 3D views without any correspondence. (b) Two partial 3D views aligned by initial feature correspondence. (c) The final fine registration between two partial 3D views.

to connect these points topologically. Another challenging issue is that there could be many holes and missing points in the point cloud. This phenomenon arises because of the low resolution of the depth images, the distant placement of cameras, and system errors.

3*D human filter*: Although partial 3D views from the RGB-D cameras have been registered into a complete 3D point cloud without distortion, it can be seen that the point cloud is still noisy and contains many isolated points and incorrectly colored points. This makes it difficult to reconstruct the point cloud as a clean polygonal mesh. Moreover, many details representing realistic effects of the user, for example, the face and clothes, must be recovered in a high quality way. To achieve this objective, we consider adopting a Gaussian filter to remove the redundant points and extract a relatively clean point cloud of a human. Fig. 9(b) shows the filtering effect on the initial point cloud in Fig. 9(a).

In addition, we plan to introduce a two-stage solution to balance these low-quality problems. In the first stage of mesh reconstruction, we initially triangulate an unorganized point cloud, and connect the massive number of points with spatial neighbors. We solve the problem of holes and burrs in the second stage, by casting the problem as a Poisson surface reconstruction. Because a global Poisson equation considers all the points simultaneously, the reconstruction is highly robust against small data noise.

*Delaunay triangulation*: During the first stage, we topologically connect and triangulate the unorganized point cloud using Delaunay triangulation [58] to obtain a coarse human mesh from the scattered point cloud. To generate an adaptive human mesh, it is possible to adjust the density of the triangles based on the number of points in different parts of a human. However, we find that Delaunay triangulation is susceptible to under-sampling and outliers. For example, there are some holes and burrs in the triangulated human mesh shown in Fig. 8(a).

*Poisson mesh reconstruction*: To overcome the problem, in the second step, we adopt Poisson mesh reconstruction [59] to refine the initial mesh, fill the holes on the surface, and remesh it. The surface reconstruction is regarded as a solution to the Poisson equation of an iso-surface given below

$$\Delta\phi = \nabla \cdot \vec{V}, \tag{6}$$

where $\phi$ is an indicator function and $\Delta$ denotes the Laplace operator. The divergence $\nabla$ of a vector field $\vec{V}$ equals Laplacian of the scalar function. The above Poisson equation is actually equivalent to expected approximation and reconstruction

$$\phi = \arg\min \| \nabla\phi - \vec{V} \|, \tag{7}$$

which means the gradient of the implicit function $\phi$ of an input point cloud $\{p_i\}$ fitting the vector field based on these points and their normal vectors $\{\vec{v}_i\}$. The form of the vector field is discretized in a subdivision octree as

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

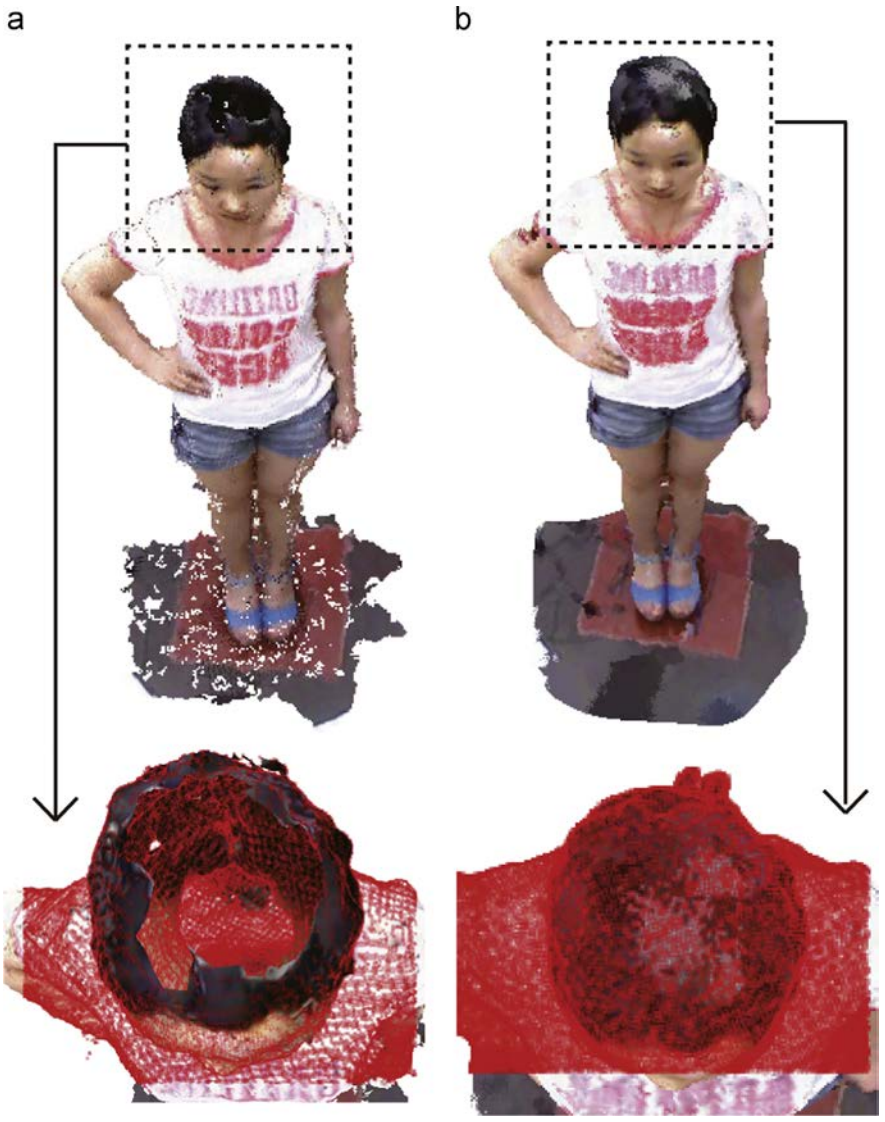


**Fig. 8.** (a) Local amplification in Delaunay triangulation. (b) Local amplification in Poisson surface reconstruction. It can be seen that holes are filled after introducing Poisson reconstruction.

follows:

$$\vec{V} = \sum_{p_i} \sum_{j \in n(p_i)} \alpha_{ij} b_j(p_i) \vec{v}_i, \tag{8}$$

where $n(p_i)$ denotes the set of the eight closest octree nodes of each point $p_i$, $j$ is the node index, and $\alpha_{ij}$ is the interpolation weight. $b_j(p_i)$ is a transformation function on each octree node $n_j$. It is worth noting that the function is similar to wavelet built on the 3D regular grid. More specifically, each transformation function is obtained by translating and scaling a fixed basis function

$$b_j(p_i) = b\left(\frac{p_i - c(n_j)}{w(n_j)}\right)\frac{1}{w(n_j)^3}, \tag{9}$$

where $b(p_i)$ denotes the basis. $c(n_j)$ is the center of the node $n_j$ and $w(n_j)$ is its width. The basis function has a uniform form, that is, the convolution with a box filter $f(x)$ (or $f(y)$, $f(z)$) $t$ times separately on each dimensional coordinate $\{x, y, z\}$. The convolution can be expressed as

$$b(x, y, z) = (f(x) * f(x))^t (f(y) * f(y))^t (f(z) * f(z))^t, \tag{10}$$

where each box filter, for example, $f(x)$, has the following form:

$$f(x) = \begin{cases} 1, & |x| < 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The approximation problem in Eq. (7) can be converted by projecting it onto the space spanned with bases $b_j$, $j \in [1, |T|]$, where $|T|$ is the number of octree nodes. The solution to $\phi$ is equivalent to minimizing

$$\sum_j \|\langle \Delta\phi - \nabla \cdot \vec{V}, b_j \rangle\|^2 = \sum_j \|\langle \Delta\phi, b_j \rangle - \langle \nabla \cdot \vec{V}, b_j \rangle\|^2, \tag{12}$$

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

which can be written in the matrix form, and the Laplace matrix $L$ with $|T| \times |T|$ elements is defined so that $L\phi$ returns the dot product of the Laplacian with each base $b_j$. Each entry of the sparse and symmetric matrix has the following expression:

$$L_{jk} = \left\langle \frac{\partial^2 b_j}{\partial x^2}, b_k \right\rangle + \left\langle \frac{\partial^2 b_j}{\partial y^2}, b_k \right\rangle + \left\langle \frac{\partial^2 b_j}{\partial z^2}, b_k \right\rangle. \quad (13)$$

Then, the Poisson equation reduces to a well-conditioned sparse linear system:

$$L\phi = \mathbf{v}, \quad (14)$$

where $\mathbf{v}$ is a $|T|-$dimensional vector whose element is as follows:

$$v_j = \langle \nabla \cdot \vec{V}, b_j \rangle. \quad (15)$$

The linear system given above is solved using a conjugate gradient solver. After obtaining the indicator function $\phi$, Marching Cubes [60] based on an octree structure are adopted to extract the iso-surface and build a 3D human mesh. The solution creates a smooth surface that robustly approximates unorganized points with missing data, as illustrated in Fig. 8(b). A comparison of the reconstruction using the initial Delaunay triangulation and Poisson surface reconstruction is shown in Fig. 9(c) and (d). In the parameter settings, the maximum depth of the octree is empirically set to 8 to balance reconstruction quality and computational effort.

*Texture mapping*: Since Poisson surface reconstruction does not allow color information to be attached to the point cloud, which is very important for recovering a realistic human, we retrieve the texture of the reconstructed mesh by searching the correspondences between points before reconstruction and vertices on the reconstructed surface. An efficient kd-tree algorithm is used to realize this task. Finally, we obtain a relatively smooth textured mesh without holes, as illustrated in Fig. 9(b).

## 7. Experimental results

*Experimental environment*: To verify the whole procedure, we built a hardware system, which includes six low-cost depth cameras (Kinect), their brackets, a dark blanket, six USB cables connecting the depth cameras to the computer, and a desktop computer with an i7 CPU and 16 GB memory. Fig. 1(a) illustrates the experimental environment, with six depth cameras placed at different locations on a circle to capture different views with as many overlapping points as possible. We do not require that these cameras are located at uniform intervals, only that they are oriented toward the human. These depth cameras simultaneously capture RGB-D images of the human. Each depth camera has been intrinsically calibrated with a checkerboard using standard calibration techniques [61,62]. Moreover, stereo calibration between the color and depth cameras of a single Kinect has been performed.

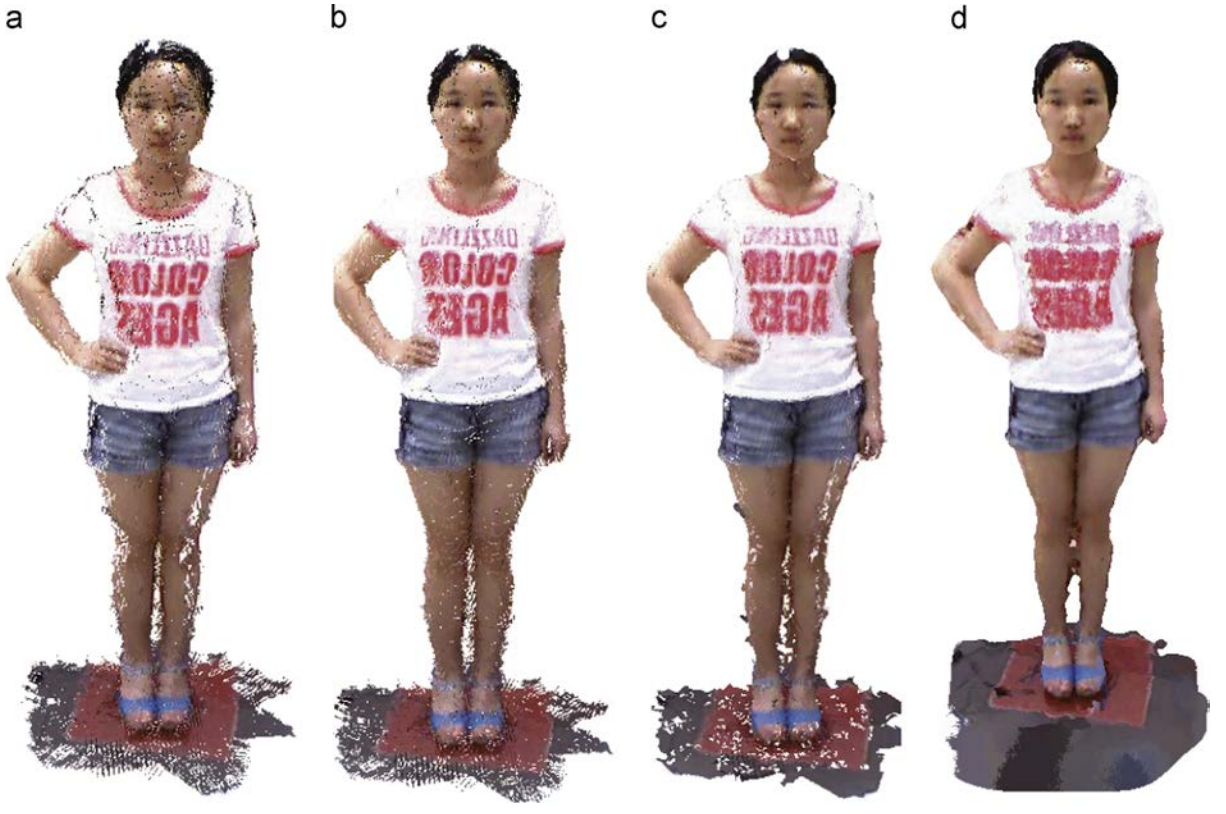


**Fig. 9.** (a) Coarse human point cloud. (b) Filtered human point cloud. (c) Delaunay triangulation. (d) Poisson surface reconstruction.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

A dark blanket is used to avoid any reflection from the floor.

3*D human reconstruction experiment*: The hardware system is used to reconstruct a real human as a color 3D mesh using the six low-quality depth cameras. A 3D mesh can be freely rendered in many applications such as games, virtual reality, and industrial human machine interaction. Fig. 1 shows the entire process from setting up the

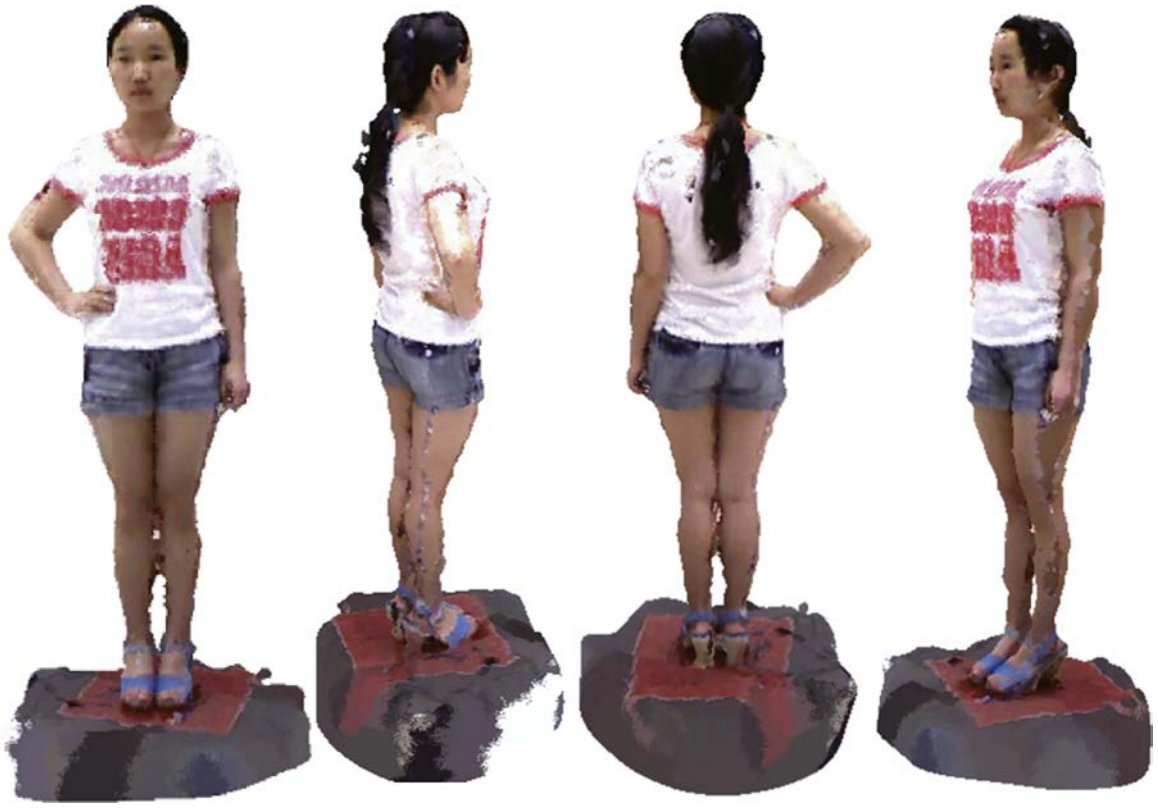

**Fig. 10.** Multi-views of a reconstructed 3D mesh model.

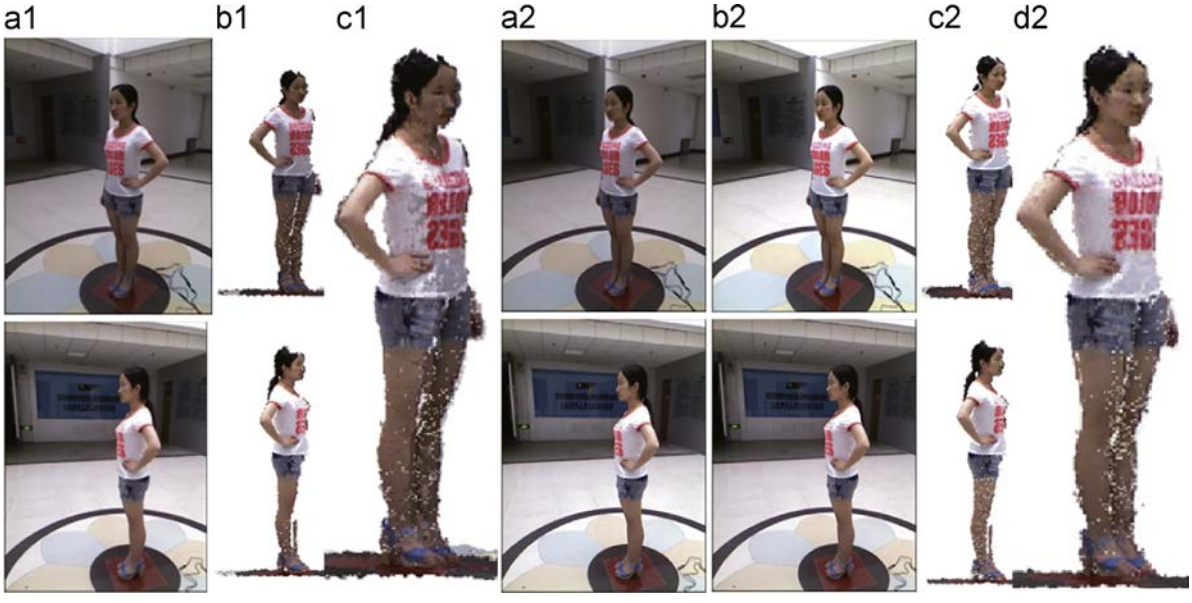


**Fig. 11.** The influence under different illumination conditions: (a1) captured images, (b1) point cloud, (c1) registration. The improved registration (d2) of point cloud (c2) by white balance (b2) from two captured images (a2).

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

hardware environment (a) to the 3D human reconstruction (f). Six RGB-D images of the human (b) are captured simultaneously. For each RGB-D image, we first remove the background to recover a 3D partial view (c) from the depth data with the help of intrinsic and stereo calibration parameters, and successively register two neighboring partial views in a 3D point cloud (d). The complete 3D point cloud (e) is globally registered using all the partial views by diffusing pairwise registration error. The final human mesh (f) is reconstructed from the 3D point

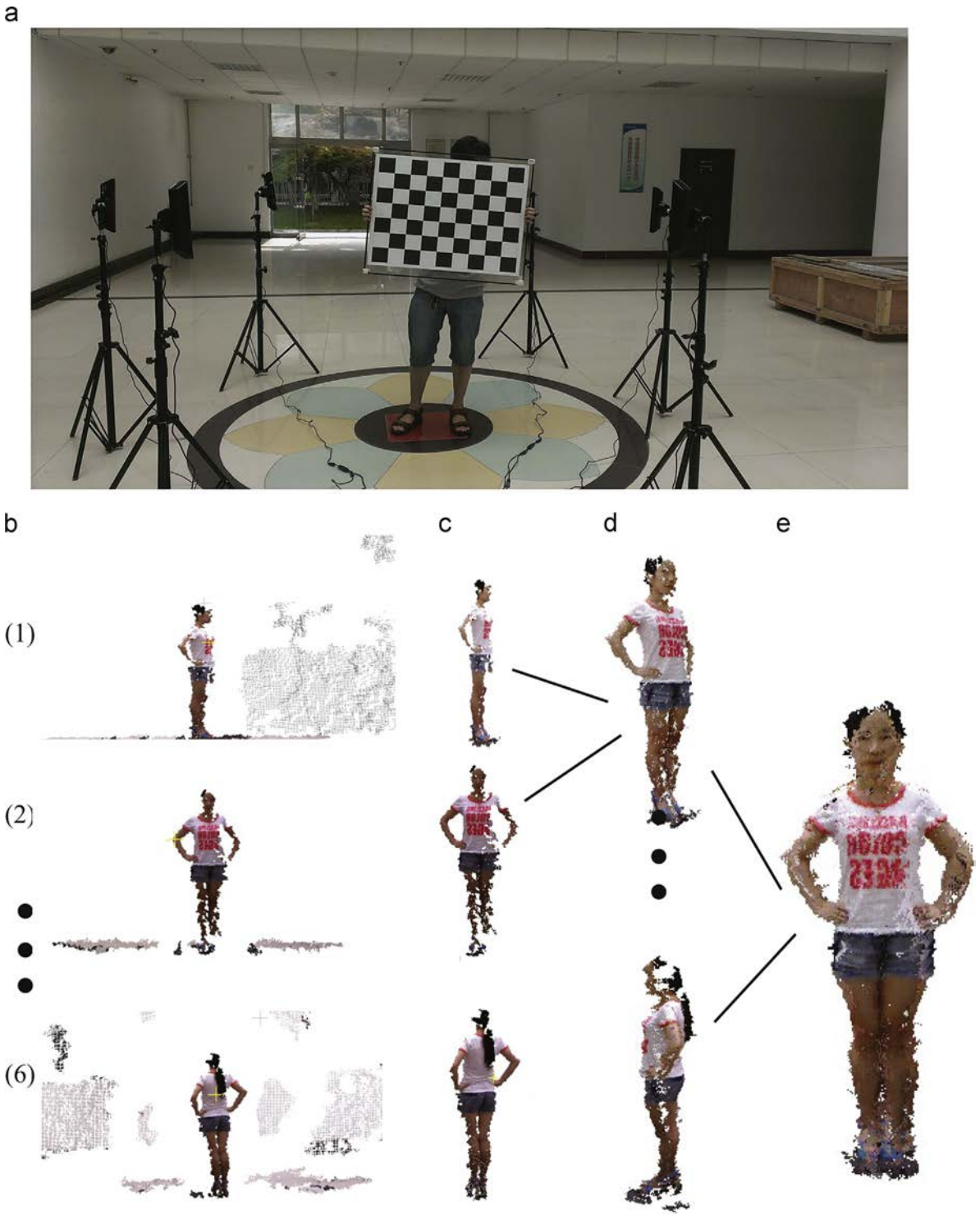


**Fig. 12.** The registration result via manual external calibration.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

cloud (e). Fig. 10 shows the reconstructed 3D mesh in detail. The mesh model can be observed from arbitrary viewpoints. We find that the reconstruction result is relatively accurate and smooth despite much noise from the depth cameras, thus demonstrating the effectiveness of our algorithm.

*Comparison with manual calibration*: We compare the automatic registration result to that based on manual external calibration of multiple cameras. Fig. 12(a) shows the hardware system used for manual calibration, where an extra large calibration board has been added. The planar checkerboard is used to provide standard corners for calibration [61], and each pair of neighboring cameras adopts the same corners. Their relative positions and calibration parameters are inferred based on the stereo projection of these corners. The six views in Fig. 12(b) were captured in our experiment, while the background in each view was removed to obtain the single color point cloud shown in Fig. 12(c) using known transformation parameters from the color camera to the depth camera of a

**Fig. 13.** The reconstruction examples of three users with different poses.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

single Kinect. Each pair of adjacent views was aligned into an overlapping color point cloud shown in Fig. 12(d) using the computed calibration parameters. All the overlapping point clouds were further integrated into a complete 3D point cloud, which can easily be reconstructed into a 3D human mesh rendered in a dynamic environment such as a home gaming environment. Fig. 12(e) shows the final manual registration result. Compared with manual calibration, our automatic algorithm achieves a better registration effect, as shown in Fig. 1(e). Moreover, our algorithm eliminates tedious manual operation, and the user does not have to hold a heavy checkerboard. This significantly improves not only the reconstruction quality but also the efficiency.

*Robustness against different illumination conditions*: While capturing the RGB-D images, we found that the registration quality is susceptible to illumination conditions. Fig. 11(a1) shows two views captured from different viewpoints. Their illumination effects differ due to the positions of the lights. Their partial 3D views are extracted individually, as shown in Fig. 11(b1). The two views are registered into an overlapping view in Fig. 11(c1). The view is clearly noisy and not smooth. We introduce an automatic white balancing method to handle this problem. The RGB color space is converted into a $YC_bC_r$ color space to detect a reference white point. A near-white region containing the reference white point is obtained by analyzing the $YC_bC_r$ coordinates. We segment each captured image into different regions and determine the reference white point. The image is automatically adjusted by using von Kries' model. For the two RGB images in Fig. 11(a2), the images after white balancing are displayed in (b2). Fig. 11 (c2) and (d2) shows the two 3D views and their registration results, respectively. It is seen that robustness against different illumination cases is enhanced and the registration result is improved.

*More reconstruction results*: We investigated the stability of our method by applying it to a group of different users with multiple poses. The group included male and female users, users with glasses, and users wearing clothes in different colors and styles. Three users with different poses were captured by our hardware system and reconstructed into 13 mesh models. Fig. 13 shows the reconstruction results, the quality of which is generally satisfactory. The average time for human reconstruction is 0.47 s. Thus, the algorithm can be implemented as an almost real-time solution for personalized avatars of users using multiple depth cameras.

*Virtual scene experiment*: To demonstrate further the realistic effect of a personalized avatar, we dynamically reconstructed a 3D human mesh for each frame in these videos, and visualized the sequential results in the game development software, Unity. We reconstructed a sequence of moving 3D human meshes with different postures

**Fig. 15.** The influence from motion blur: an unsuccessful reconstruction model.


**Fig. 14.** (a) Reconstructed dancing action. (b) Reconstructed walking action. The reconstructed realistic 3D human is placed in a virtual scene, and follows user movements.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

to test our method, as shown in Fig. 14. These 3D human meshes were dynamically imported into a virtual scene built in Unity, and composed a realistic and meaningful action, which approximates actual human movements in the scene of a game or virtual reality. Using this method, the user's interactive experience can be enhanced. It should be noted that in practical applications the reconstruction of a realistic 3D human mesh can be performed only once, for example, in the system initialization stage. The movement of a skeleton can be mapped onto a preliminary 3D human in real time.

## 8. Conclusion

In this paper, we presented a novel proposal for 3D real human reconstruction of real users by employing multiple depth cameras. The core of the implementation process lies in how to reconstruct realistic humans from noisy RGB data. The process consists of three key steps: 3D partial view generation, registration of partial views, and human mesh reconstruction. The novel application allows a user to be truly immersed in a virtual really scene using low cost RGB-D cameras. This should greatly enhance the interactive experience in a virtual world.

*Limitations*: Our approach suffers from several limitations. The first is that we are unable to solve the fact that the reconstruction is readily affected by instantaneous motion blur. Users are always in motion. Sometimes their gestures change with high frequency, which exceeds the capturing ability of the low-cost cameras. Such quick motions cannot be followed by a camera. As a result, some depth values of the moving parts are lost, which results in reconstruction failure, as shown in Fig. 15. The defect appears in the user's right arm denoted by the dotted circle. We also observed another failure mode. If the overlapping region between two views is almost flat, it is difficult to achieve a reliable accurate pairwise registration. Since the features in the overlap are similar, it is difficult to find the correct correspondences for two views. Fig. 16 shows an example. Two different views (a) are registered from an initial alignment (b) into the result (c). We see that the final registration has obvious defects.

In the future, we intend to concentrate on improving the quality of the reconstructed human mesh from sparse RGB-D data in cluttered environments. Neighboring frames in RGB-D videos and their distributed correlations [63] may be considered as complementary information, and different reconstructed humans in each frame can be nonrigidly deformed into a more complete

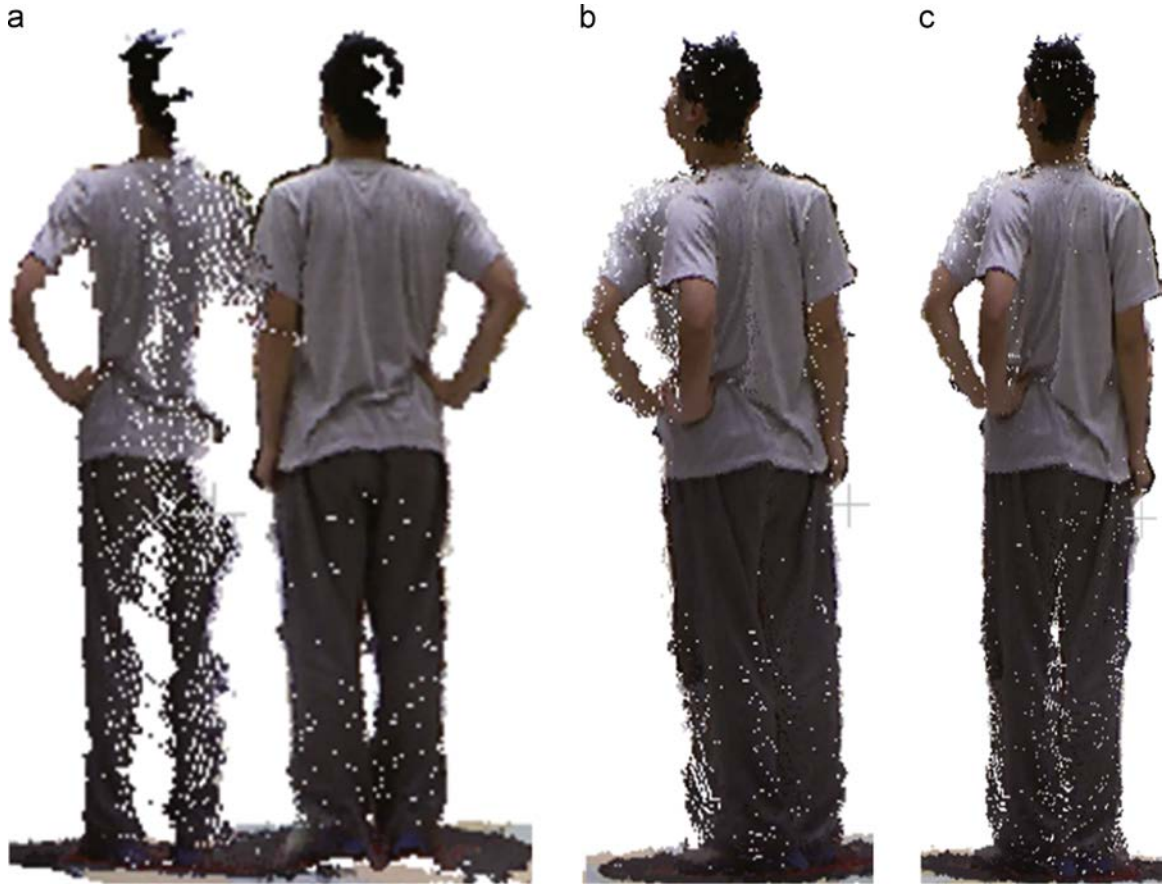


**Fig. 16.** The failure case of registration: (a) two different views, (b) the initial alignment, and (c) the defective registration.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

3D mesh. Moreover, we intend solving practical applications such as a virtual fitting room to validate further the performance of our approach.

## Acknowledgments

The work is partially supported by NSFC (Nos. 61003137, 61202185, 61473231, and 91120005), NWPU Basic Research Fund 310201401 (JCQ01009 and JCQ01012), Fund of National Engineering Center for Commercial Aircraft Manufacturing (201410), National Aerospace Science Foundation of China (2014ZD53), and Open Fund of State Key Lab of CAD & CG in Zhejiang University (A15).

## References

[1] M. Berger, J.A. Levine, L.G. Nonato, G. Taubin, C.T. Silva, A benchmark for surface reconstruction, ACM Trans. Graph. 32 (2) (2013) 20:1–20:17.
[2] Microsoft Kinect, ⟨http://www.xbox.com/kinect⟩, 2010.
[3] D.S. Alexiadis, P. Kelly, P. Daras, N.E. O'Connor, T. Boubekeur, M.B. Moussa, Evaluating a dancer's performance using kinect-based skeleton tracking, in: Proceedings of the ACM Multimedia, 2011, pp. 659–662.
[4] Z. Liu, S. Tang, H. Qin, S. Bu, Evaluating user's energy consumption using kinect based skeleton tracking, in: Proceedings of ACM Multimedia, 2012, pp. 1373–1374.
[5] A. Bleiweiss, D. Eshar, G. Kutliroff, A. Lerner, Y. Oshrat, Y. Yanai, Enhanced interactive gaming by blending full-body tracking and gesture animation, in: Proceedings of the ACM SIGGRAPH Asia Sketches, 2010.
[6] F. Pedersoli, N. Adami, S. Benini, R. Leonardi, Xkin-extendable hand pose and gesture recognition library for kinect, in: Proceedings of the ACM Multimedia, 2012, pp. 1465–1468.
[7] J. Tong, J. Zhou, L. Liu, Z. Pan, H. Yan, Scanning 3D full human bodies using kinects, IEEE Trans. Vis. Comput. Graph. 18 (4) (2012) 643–650.
[8] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, Scape: shape completion and animation of people, ACM Trans. Graph. 24 (3) (2005) 408–416.
[9] A. Weiss, D. Hirshberg, M.J. Black, Home 3D body scans from noisy image and range data, in: Proceedings of the International Conference on Computer Vision, 2011, pp. 1951–1958.
[10] Y. Chen, Z. Liu, Z. Zhang, Tensor-based human body modeling, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 105–112.
[11] D.S. Alexiadis, D. Zarpalas, P. Daras, Real-time, full 3-D reconstruction of moving foreground objects from multiple consumer depth cameras, IEEE Trans. Multimed. 15 (2) (2013) 339–358.
[12] T. Helten, M. Müller, H.-P. Seidel, C. Theobalt, Real-time body tracking with one depth camera and inertial sensors, in: Proceedings of the International Conference on Computer Vision, 2013, pp. 1105–1112.
[13] O. Gedik, A. Alatan, 3-D rigid body tracking using vision and depth sensors, IEEE Trans. Cybern. 43 (5) (2013) 1395–1405.
[14] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 1297–1304.
[15] M. Wang, R. Hong, X.-T. Yuan, S. Yan, T.-S. Chua, Movie2comics: towards a lively video content presentation, IEEE Trans. Multimed. 14 (3) (2012) 858–870.
[16] B. Leng, Z. Xiong, Modelseek: an effective 3D model retrieval system, Multimed. Tools Appl. 51 (3) (2011) 935–962.
[17] Z. Liu, S. Bu, K. Zhou, S. Gao, J. Han, J. Wu, A survey on partial retrieval of 3D shapes, J. Comput. Sci. Technol. 28 (5) (2013) 836–851.
[18] Y. Gao, Q. Dai, View-based 3D object retrieval: challenges and approaches, IEEE MultiMed. 21 (3) (2014) 52–57.
[19] A.M. Bronstein, M.M. Bronstein, L.J. Guibas, M. Ovsjanikov, Shape google: geometric words and expressions for invariant shape retrieval, ACM Trans. Graph. 30 (1) (2011) 1–20.
[20] B. Leng, X. Zhang, M. Yao, Z. Xiong, 3D object classification using deep belief networks, in: MultiMedia Modeling, Lecture Notes in Computer Science, vol. 8326, 2014, pp. 128–139.
[21] B. Leng, X. Zhang, M. Yao, X. Zhang, A 3D model recognition mechanism based on deep Boltzmann machines, Neurocomputing, 2014, http://dx.doi.org/10.1016/j.neucom.2014.06.084, in press.
[22] R. Ji, H. Yao, X. Sun, B. Zhong, W. Gao, Towards semantic embedding in visual vocabulary, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 918–925.
[23] V. Barra, S. Biasotti, 3D shape retrieval using kernels on extended Reeb graphs, Pattern Recognit. 46 (11) (2013) 2985–2999.
[24] Z. Lian, A. Godil, B. Bustos, M. Daoudi, J. Hermans, S. Kawamura, Y. Kurita, G. Lavoué, H.V. Nguyen, R. Ohbuchi, Y. Ohkita, Y. Ohishi, F. Porikli, M. Reuter, I. Sipiran, D. Smeets, P. Suetens, H. Tabia, D. Vandermeulen, A comparison of methods for non-rigid 3D shape retrieval, Pattern Recognit. 46 (1) (2013) 449–461.
[25] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, M. Alexa, Sketch-based shape retrieval, ACM Trans. Graph. 31 (4) (2012) 31:1–31:10.
[26] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, N. Zhang, Less is more: efficient 3D object retrieval with query view selection, IEEE Trans. Multimed. 11 (5) (2011) 1007–1018.
[27] Y.-S. Liu, K. Ramani, M. Liu, Computing the inner distances of volumetric models for articulated shape description with a visibility graph, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2538–2544.
[28] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, T.-S. Chua, Camera constraint-free view-based 3D object retrieval, IEEE Trans. Image Process. 21 (4) (2012) 2269–2281.
[29] M. Wang, Y. Gao, K. Lu, Y. Rui, View-based discriminative probabilistic modeling for 3D object retrieval and recognition, IEEE Trans. Image Process. 22 (4) (2013) 1395–1407.
[30] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai, 3-D object retrieval and recognition with hypergraph analysis, IEEE Trans. Image Process. 21 (9) (2012) 4290–4303.
[31] R. Ji, X. Xie, H. Yao, W.-Y. Ma, Mining city landmarks from blogs by graph modeling, in: Proceedings of the ACM Multimedia, 2009, pp. 105–114.
[32] B. Leng, Z. Qin, A powerful relevance feedback mechanism for content-based 3D model retrieval, Multimed. Tools Appl. 40 (1) (2008) 135–150.
[33] C.-H. Shen, H. Fu, K. Chen, S.-M. Hu, Structure recovery by part assembly, ACM Trans. Graph. 31 (6) (2012) 180:1–180:11.
[34] Y.M. Kim, N.J. Mitra, D.-M. Yan, L. Guibas, Acquiring 3D indoor environments with variability and repetition, ACM Trans. Graph. 31 (6) (2012) 138:1–138:11.
[35] I. Kakadiaris, D. Metaxas, Three-dimensional human body model acquisition from multiple views, Int. J. Comput. Vis. 30 (3) (1998) 191–218, http://dx.doi.org/10.1023/A:1008071332753.
[36] C.C. Wang, Parameterization and parametric design of mannequins, Comput. Aided Des. 37 (1) (2005) 83–98.
[37] J.-W. Hsieh, C.-H. Chuang, S.-Y. Chen, C.-C. Chen, K.-C. Fan, Segmentation of human body parts using deformable triangulation, IEEE Trans. Syst Man Cybern. Part A: Syst. Humans 40 (3) (2010) 596–610.
[38] M. Holte, C. Tran, M. Trivedi, T. Moeslund, Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments, IEEE J. Select. Top. Signal Process. 6 (5) (2012) 538–552.
[39] N. Werghi, Segmentation and modeling of full human body shape from 3-D scan data: a survey, IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. 37 (6) (2007) 1122–1136.
[40] S. Chris, W. Grimson, Adaptive background mixture models for real-time tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 1999, p. 252.
[41] K. Lai, L. Bo, X. Ren, D. Fox, A large-scale hierarchical multi-view rgb-d object dataset, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2011, pp. 1817–1824.
[42] B. Leng, S. Guo, X. Zhang, X. Zhang, 3D object retrieval with stacked local convolutional autoencoder, Signal Process. (2014), in this issue.
[43] J. Zeng, B. Leng, X. Zhang, 3-D object retrieval using topic model, Multimed. Tools Appl. (2014), in press.
[44] Y. Gao, M. Wang, R. Ji, X. Wu, Q. Dai, 3-D object retrieval with Hausdorff distance learning, IEEE Trans. Ind. Electron. 61 (4) (2014) 2088–2098.
[45] S. Berretti, A.D. Bimbo, P. Pala, Partial match of 3D faces using facial curves between SIFT keypoints, in: Proceedings of the Eurographics Workshop on 3D Object Retrieval, 2011, pp. 117–120.
[46] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[47] Y. Gao, Q. Dai, N.-Y. Zhang, 3D model comparison using spatial structure circular descriptor, Pattern Recognit. 43 (3) (2010) 1142–1151.
[48] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, W. Gao, Location discriminative vocabulary coding for mobile landmark search, Int. J. Comput. Vis. 96 (3) (2012) 290–314.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021

[49] R. Ji, H. Yao, W. Liu, X. Sun, Q. Tian, Task-dependent visual-codebook compression, IEEE Trans. Image Process. 21 (4) (2012) 2282–2293.
[50] F. Tombari, S. Salti, L. D. Stefano, Unique signatures of histograms for local surface description, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 356–369.
[51] F. Tombari, S. Salti, L.D. Stefano, A combined texture-shape descriptor for enhanced 3D feature matching, in: Proceedings of the IEEE International Conference on Image Processing, 2011, pp. 809–812.
[52] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009) 733–746.
[53] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Trans. Image Process. 21 (11) (2012) 4649–4661.
[54] P. Li, M. Wang, J. Cheng, C. Xu, H. Lu, Spectral hashing with semantically consistent graph for image indexing, IEEE Trans. Multimed. 15 (1) (2013) 141–152.
[55] M. Leordeanu, M. Hebert, A spectral technique for correspondence problems using pairwise constraints, in: Proceedings of the International Conference on Computer Vision, vol. 2, 2005, pp. 1482–1489.
[56] V. Kolmogorov, C. Rother, Minimizing nonsubmodular functions with graph cuts-a review, IEEE Trans. Pattern Anal. Mach. Intell. 29 (7) (2007) 1274–1279.
[57] A. Myronenko, X. Song, Point set registration: coherent point drift, IEEE Trans. Pattern Anal. Mach. Intell. 32 (12) (2010) 2262–2275.
[58] P. Cignoni, C. Montani, R. Scopigno, Dewall: a fast divide and conquer Delaunay triangulation algorithm in Ed, Comput. Aided Des. 30 (5) (1998) 333–341.
[59] M. Kazhdan, M. Bolitho, H. Hoppe, Poisson surface reconstruction, in: Proceedings of the Eurographics Symposium on Geometry Processing, 2006, pp. 61–70.
[60] W.E. Lorensen, H.E. Cline, Marching cubes: a high resolution 3D surface reconstruction algorithm, in: Proceedings of the ACM SIGGRAPH, 1987, pp. 163–169.
[61] Z. Zhang, A flexible new technique for camera calibration, IEEE Trans. Pattern Anal. Mach. Intell. 22 (11) (2000) 1330–1334.
[62] H.C. Daniel, K. Juho, H. Janne, Joint depth and color camera calibration with distortion correction, IEEE Trans. Pattern Anal. Mach. Intell. 34 (10) (2012) 2058–2064.
[63] R. Ji, L.-Y. Duan, J. Chen, L. Xie, H. Yao, W. Gao, Learning to distribute vocabulary indexing for scalable visual search, IEEE Trans. Multimed. 15 (1) (2013) 153–166.

Please cite this article as: Z. Liu, et al., 3D real human reconstruction via multiple low-cost depth cameras, Signal Processing (2014), http://dx.doi.org/10.1016/j.sigpro.2014.10.021