High-precision Human Body Acquisition via Multi-Binocular Stereopsis

Qing Ran [1], Kaimao Zhou [1], Yong-Liang Yang [2], Junpeng Kang [1], Linan Zhu [1], Yizhi Tang [1], Jieqing Feng [1*]

[1] State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, P.R. China
[2] University of Bath, Bath, BA2 7AY, United Kingdom
* Author for correspondence (jqfeng@cad.zju.edu.cn)

Abstract

It remains challenging how to acquire a human body shape with high precision and evaluate the reconstructed models effectively, because the results can be easily affected by various factors (e.g., the performance of the capture device, the unwanted movement of the subject, and the self-occlusion of the articulated body structure). To tackle the above challenges, this research presents a passive acquisition system, which comprises 60 spatially-configured Digital Single Lens Reflex (DSLR) cameras and a carefully devised algorithmic pipeline for shape acquisition in a single shot. Different from traditional multi-view stereo solutions, the constituent cameras are synchronized and organized into 30 binocular stereo rigs to capture images from multiple views simultaneously. Each binocular stereo rig is regarded as a depth sensor. The acquisition pipeline consists of three stages. First, camera calibration is performed to estimate intrinsic and extrinsic parameters of all cameras, especially for paired binocular cameras. Second, depth inference based on stereo matching is employed to recover reliable depth information from RGB images. A novel hierarchical seed-propagation stereo matching framework is proposed, resulting in 30 dense and uniform-distributed partial point clouds. Finally, a point-based geometry processing step composed of multi-view registration and surface meshing is carried out to obtain high-quality watertight human body shapes. This research also proposes an elaborate and novel method to assess the accuracy of reconstructed non-rigid human body model based on anthropometry parameters, which solves the synchronization of the ground-truth values and the measured values. Experimental results show that the system can achieve the reconstruction accuracy within 2.5 millimeters in average.

1 Introduction

Interests in acquiring high-precision 3D human body shapes are motivated by a wide range of applications, such as medical rehabilitation, garment customization, virtual fitting, etc. This task is significantly challenging, primarily because the human body is non-rigid which easily varies during the acquisition process. Moreover, how to evaluate the reconstructed human body models is rather difficult due to the problem of synchronization between the ground-truth values and the measured values. For a fair comparison, the ground-truth should be simultaneously obtained during shape acquisition and then compared with the acquired shape. Thus, this research attempts to address the above two problems via a passive multi-binocular vision system with synchronized DSLR cameras, especially the issue of accuracy evaluation.

A growing body of literature has been examined in the field of human body model acquisition. The most common approach is to use expensive high-end active devices, such as 3D scanners based on laser ranging or industrial structure light, which could result in detailed human body point clouds. Due to self-occlusion and limited scanning range, the capturing could not be instantaneous. It will lead to shape and texture distortions due to even a small movement of human body when conducting multi-view captures. Various geometry processing algorithms are proposed to estimate the non-rigid deformation and then integrate all scanned point clouds into a complete human body shape. Several high precision human body datasets from [4, 5, 17] have been collected in this way with considerable costs and play important roles in subsequent research, such as model analysis. Recent works from [49, 29, 8, 54] mainly focus on low-cost, portable consumer depth sensors such as Kinect or RGB-D cameras. Because of the lower resolution of the depth images provided by those sensors, the obtained human body shape may lack geometric features, even if prior knowledge such as a detailed parametric template has been provided. Currently, consumer depth sensors based acquisition method is more suitable for motion capture or human body tracking.

One thing that should be addressed in active approaches is that, either high-end or low-cost systems take at least several seconds to scan a complete human body. This relatively longer process is mainly due to the limited scanning efficiency of high-end devices (e.g., laser scanners), and the interference between active sensors from different views (especially for structured light based sensors). Therefore, the geometry and texture information of a human body can hardly be obtained instantly and simultaneously, which is a key technical challenge that results in shape and texture distortions caused by body movement.

Passive approaches utilize techniques of image-based modeling as proposed by [33, 44] to solve the challenge of capturing time. With no constraints on the arrangement and number of cameras, a human body could be captured in a single shot, approximate to instantaneous. Then the human body shape along with textures is reconstructed based on photometric geometry. However, due to the restricted image quality of the capture devices, the reliability and accuracy of passive approaches are traditionally regarded as being inferior to active methods. Currently, high resolution cameras (such as DSLR) are able to capture rich geometry and texture details of human body, which benefit recovering geometry of the human body surface from images. With the development in multi-view stereopsis (cf. [37, 38]), there is a great potential for passive approaches to be comparable with the active methods in terms of human body acquisition performance.

This research proposes a carefully-designed passive full-body capture system which consists of multiple synchronized DSLR cameras, to acquire high precision models of static human body. In the proposed system, the shape distortion can be significantly reduced, and both geometry and texture can be obtained simultaneously. To take full advantage of high-resolution images in every viewpoint, this study employs the depth map fusion method [40] to reconstruct a human body model. Instead of computing dense point clouds through multi-view stereopsis, a binocular stereo rig is used as a depth sensor to generate a dense depth map for each viewpoint by performing stereo matching. It results in a key challenge that how to robustly and effectively estimate dense depth information of high resolution stereo images of human body, which is texture-less in general. Another key challenge is to integrate all the partial point clouds into a complete human body model, which requires highly accurate estimation of global extrinsic parameters of each camera. The evaluation of reconstructed models remains challenging because it is difficult to obtain the ground-truth value and the measured values simultaneously. All methods of using off-line measurements or scans are not fair enough because the human body shape changes all the time. In this paper, the above key challenges, such as weak texture

of human skin in stereo matching, and measurement of acquired models, etc., are carefully investigated and tackled. The main contributions of this article include:

- A high precision 3D models of static human body acquisition and reconstruction system is designed and developed, including specified hardware configuration and detailed acquisition algorithmic pipeline.
- A hierarchical stereo matching method based on seed-propagation is proposed to robustly estimate the depth information of high resolution stereo images of human body, which is texture-less in general.
- An elaborate and novel method to assess the accuracy of reconstructed non-rigid human body model based on isometric geometry and congruent constraints via anthropometry parameters. In the proposed evaluation method, the measured values and the ground-truth values could be obtained simultaneously.

Comprehensive experimental results will be presented to verify the performance of the proposed system.

2 Related Work

Previous human body acquisition works based on passive-vision will be reviewed since they are most relevant to our work. These works are briefly categorized into template-based and template-free methods according to whether a template prior is used for reconstruction.

Template-based methods fit a pre-defined template model to partial or insufficient point clouds, so as to acquire completed models. Key problems in template fitting, including vertex corresponding, hole filling and surface meshing, are solved effectively in [4]. Following this pioneering work, many template-based reconstruction methods have been developed. A generic model of human shape and kinematic structure are optimized in [41] to simultaneously match stereo, silhouette, and feature data across multiple views. The naked human body shape under clothing is estimated in [55] by fitting a parametric model to 3D scans. Pre-defined templates could be fitted to images directly, which is common in lightweight modeling application. In [6], detailed human body are achieved by estimating the fitting parameters of the SCAPE models [5] directly from images. Guan et al. [15] acquire both the shape and pose from a single photograph using a set of markers on the SCAPE model specified by users.

Template-based methods can efficiently generate a complete human body model with no hole, even with less captured data from several views. However, the result quality may be limited by the shape representation ability of the pre-defined template. For example, the template defined in a low-dimensional shape space may filter out high-frequency geometric details of a human body. Furthermore, the fixed parameterization of the template may not be able to capture human body shape variations, especially with topology changes.

Template-free methods utilize multi-view stereopsis (MVS) based reconstruction to obtain a complete model (cf. [40]). The first MVS framework is proposed by [43] for modeling urban scenes and general objects. Many assumptions (such as planar primitives) play an important role in effectively recovering the surface geometry information [53, 16]. Furukawa et al. [14] achieves quasi-dense 3D reconstruction by recovering a number of small rectangular patches covering the object visible in the images, which known as the patch-based multi-view stereopsis algorithm (PMVS). High resolution images make it possible to reconstruct dense geometry directly. A dense 3D environment modeling method is proposed in [23] by using multiple pairs of high-resolution spherical images. The accurate multi-view reconstruction method [9] exploits the high resolution images to acquire static models of indoor scene/objects. Its multi-binocular stereo pipeline is similar to the proposed method. However, due to the simple depth fusion strategy based on the visibility, alignment errors may exist in the eventually completed point cloud in [9].

The flexibility of MVS based reconstruction method extends its applicability to the domain of accurate static human modeling in recent decades. Human faces as in [7, 52] and human bodies as in [34, 3] have been captured by many acquisition systems based on MVS. The surface of human body is recovered from a video via robust stereo matching in [30]. It makes use of many texture-related information such as visual hull, frontier points and implicit points to boost surface completeness and accuracy. However, as a natural weakness, the human body skin lacks of textures. Other researches [42, 27], which aim at the capture of dynamic human motion, prefer to replace the human body templates in conventional methods with data of surface recovered by MVS. The limited number of cameras in the motion capturing may lead to the existence of wide-baseline stereo. DAISY [48] designs an effective feature which could eliminate large distortions in that case, and then estimate dense depth maps from stereo image pairs. The temporal information [26] in captured videos could be also used to refine the MVS reconstruction. Tung et al. [51] take advantage of the image content stability provided by each single-view video to recover any surface region visible by at least one camera. Recent work [12] focuses on transforming free-viewpoint video from multi-modal images, including RGB images, IR images, etc. High quality human body models could be reconstructed by combining comprehensive information.

Similar to this paper, Remondino [34] investigates the reconstruction of static human body shapes from un-calibrated image sequences. They focus on the estimation of camera orientation, and the average error of the reconstructed model is about $6.0mm$ by rough manual measurements. Beeler et al. [7] present an impressive multi-view system which consists of multiple expensive full-frame cameras for 3D human face capture. Images were taken at a close range so that the pores of the face can be used as features for stereo matching and microscopic geometry recovery. The reconstruction error is estimated for a physical mask, which is not directly applicable to real faces because the surface reflectance of the face mask is different from that of human skin. To obtain 3D photo-realistic virtual avatars for just-in-time use in a game or simulation, Feng et al. [13] capture the human face, hands and the whole body separately and then stitched together. Joo et al. [21] present an approach to capture the 3D structure and motion of a ground of people, which extents the application of the MVS-based capture system.

3 System Overview

Inspired by [7], this article presents a passive multi-binocular system to capture a static human body in a single shot, aiming at the accuracy of anthropometry measurements of the reconstructed human body. In this section, an overview of the proposed system is described, which includes the hardware configuration, and the acquisition algorithmic pipeline which turning raw captured images into high-quality human body model.

3.1 Hardware Overview

The system hardware configuration is shown in Fig. 1. 60 DSLR cameras (Canon 600D) with $50mm$ fixed-focal lens are placed around a circular capture space of which diameter is $5m$. All the cameras are arranged into 30 meta units. Each meta unit is a stereo rig of two cameras with an accurate baseline $180mm$. Among all the stereo rigs, 24 of them are evenly distributed along the circle from 8 circular angles, focusing on the main torso of a human body from top, middle and bottom views for each angle.

Figure 1: The multi-view system setup with binocular stereo rigs.

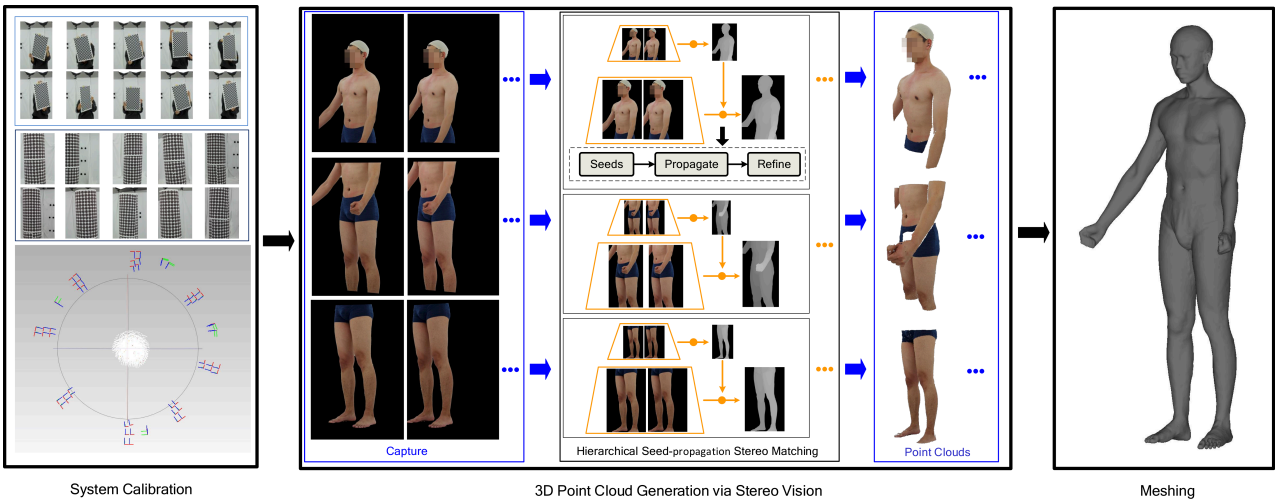


Figure 2: Overview of the acquisition pipeline with three major stages.

Other 6 stereo rigs are arranged for two arms with high flexibility. Each arm is captured from 3 viewpoints, including front-view, back-view and side-view, to ensure flexible pose space for arms during capturing. The proposed hardware setup guarantees redundant overlaps between adjacent viewpoints and covers the captures of various body shapes and heights (up to $2.0m$)

Each camera is connected with a wireless shutter. All shutters can be triggered by the same remote controller. In this way, all cameras are synchronized with a error of 0.5 milliseconds so that raw image data could be captured almost simultaneously. The resolution of captured images is 5148×3456 and pixels of the human body could account for more than 50%. Instead of commonly used green background, the proposed system uses white background to eliminate the color interference between background and human skin. A diffuse environment lighting is set by using several photography lamps. Light first arrives at white ceiling and then reflects to human skin which prevents specular highlights. It should be noticed that the number of cameras in the proposed system could be adjusted according to capture requirements.

3.2 Acquisition Overview

As shown in Fig. 2, taking a set of uncalibrated images captured from multiple views as input, there are three stages in the acquisition pipeline: system calibration, depth recovery, and 3D surface reconstruction. The first stage is to estimate poses of all cameras in the global coordinate system. A checkerboard pattern and a cylindrically distributed pattern are used for calibrating each stereo rig and all cameras respectively. The second stage is depth recovery via binocular stereo vision. A novel hierarchal seed-propagation stereo matching framework is proposed to generate a dense and accurate depth map for each stereo rig. Apart from typical stereo vision refinement, this study seeks further by applying 3D geometric refinement techniques to obtain smoother depth maps. The third stage is to generate a complete mesh from multiple partial point clouds recovered in the previous stage. Point-based processing pipeline, including multi-view point clouds registration, hole filling and 3D reconstruction, is employed to reconstruct a complete human body model. The accuracy of system calibration and depth recovery ensures the quality of the reconstructed human body models. The next section elaborates the details of the acquisition pipeline.

Recently, deep learning based methods have been extensively studied. Learning-based refinement strategies [27] are used to benefit the reconstruction of arbitrary shapes. An end-to-end learning framework for multi-view stereopsis is proposed in [20]. There are several key challenges when applying the learning-based techniques, such as the ground-truth of camera parameters and the human body models in our capture system, proper loss function which is effective in estimating the human body surface, etc. Note that all previous works either showed only qualitative results, or roughly measured the results which exhibit larger reconstruction error.

4 Human Body Model Acquisition

4.1 System Calibration

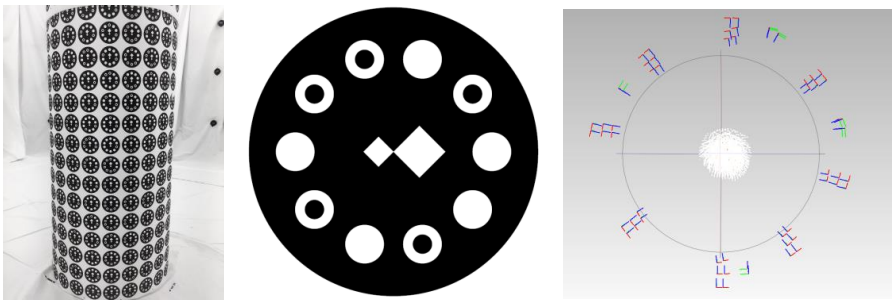
This stage includes local calibration for each stereo rig and global calibration for the system, aiming to estimate the intrinsic and extrinsic parameters of all cameras. To improve the accuracy of system calibration, a 3D calibration object (as shown in Fig.3) is designed, which feeds accurate matched features in the optimization of Bundle Adjustment (BA). Meanwhile, the BA algorithm is augmented with constraints of relative external parameters within a stereo rig. Details are described as follows.

For each stereo rig with two cameras $\{C_i^l, C_i^r\}$, this study estimates intrinsic parameters $\{K_i^l, K_i^r\}$ and stereo extrinsic parameters $\{R_i, t_i\}$ by using a checkerboard calibration pattern as in [56]. The results $\{K_i^l, K_i^r, R_i, t_i\}$ will be used as optimization constraints to reduce uncertainties in the following global calibration. Note that local calibration only needs to be performed once, since each stereo rig is formed by firmly mounting two cameras on a horizontal/vertical gimbal.

Global calibration estimates the projection matrices for all cameras in a global coordinate system, benefiting the final multi-view depth fusion. The local calibration result is introduced as a constraint into the objective function of bundle adjustment [50]. An optimal 3D structure $X = \{X_j\}$ and viewing parameters $C = \{K_i, P_i\}$ are solved by minimizing Eq. (1) in the bundle adjustment.

$$G(X, C) = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} ||q_{ij} - K_i \cdot P_i \cdot X_j||^2. \tag{1}$$

Here K_i, P_i indicate the intrinsic matrix of the i-th camera C_i, and the estimated projection matrix (i.e., the camera pose). X_j is the estimated j-th 3D feature in the scene, while q_{ij} indicates the corresponding 2D matched feature in the image of the i-th camera. w_{ij} is an indicator variable which represents the visibility of X_j in C_i. n and m are the number of cameras and matched 2D features $\{q_{ij}\}$, respectively. In practice, the optimization of Eq. (1) may fail due to too many unknown parameters.



(a) (b) (c)

Figure 3: (a) The global calibration cylinder. (b) One example of the encoding patterns. (c) The system calibration result. Each coordinate frame represents a camera. The points in the middle represent the reconstructed 3D encoding points on the cylinder.

This study adopts two strategies to reduce uncertainties. First, a 3D calibration cylinder shown in Fig. 3(a) is designed to collect sufficient and reliable matched 2D features - $\{q_{ij}\}$. Each printed pattern is coded as a unique feature point to ensure accurate matching of $\{q_{ij}\}$. As shown in Fig. 3(b), small solid and hollow disks represent 1 and 0. The disk pointed by the two aligned squares in the middle is the starting point of the code. Second, the local calibration result is introduced into the objective function as follows:

$$\bar{G}(C, X) = G(C, X) + \gamma \cdot \sum_{i=1}^{s} ||P_i^L - [R_i|t_i] \cdot {P_i}^R||_F^2 \tag{2}$$

Compared with Eq. (1), the additional term in Eq. (2) constrains the estimated mutual camera poses in a stereo rig to be consistent with the stereo extrinsic parameters from the local calibration. Moreover, $\{K_i\}$ are also given by the local calibration. The encoding matched feature points $\{q_{ij}\}$ are provided accurately by the 3D cylinder. γ is a scalar to adjust the weight of stereo extrinsic constraint and set to 1 in the practical experiments. Levenberg-Marquardt minimization [32] is applied to solve the optimal camera poses. Fig. 3(c) shows the calibration results. Stereo rigs can be clearly observed and are consistent with the camera arrangement in Fig. 1.

4.2 Hierarchical Seed-Propagation Stereo Matching

To recover the depth information from each stereo view, this study estimates a disparity image from a pair of stereo images via stereo matching. However, the lack of colored textures on human skin may lead to enormous matching ambiguities. To tackle this problem, a hierarchical stereo matching method based on seed-propagation [28] is proposed to robustly estimate the depth information of high resolution stereo images of human body. Its pipeline is shown as Fig. 4,

First, an image pyramid is established to accelerate the efficiency of stereo matching of high resolution images. In each level of the image pyramid, an interaction-between-levels stereo matching algorithm via seed propagation is proposed, based on two observations: the smoothness of human skin and the existence of local salient features therein. Moreover, two 3D point cloud processing operations are employed to optimize the 2D disparity map, so that the recovered depth is consistent with the human shape geometry prior as much as possible. In following, the overall hierarchical strategy is presented first, then details of individual steps at a single level of the hierarchy are elaborated.

4.2.1 Hierarchical Stereo Matching Framework

Image pre-processing is initially performed on exported raw images from all cameras, resulting in high-resolution images (about 3000×4000) used for stereo matching. The image pre-processing includes RAW-to-RGB format converting, background cropping and rectification. Particularly, during the RAW-to-RGB format converting, several pairs of stereo images are generated in each view by applying different photometric rendering parameters to the raw image data (i.e., CR2 format image data). Thus, radiometric

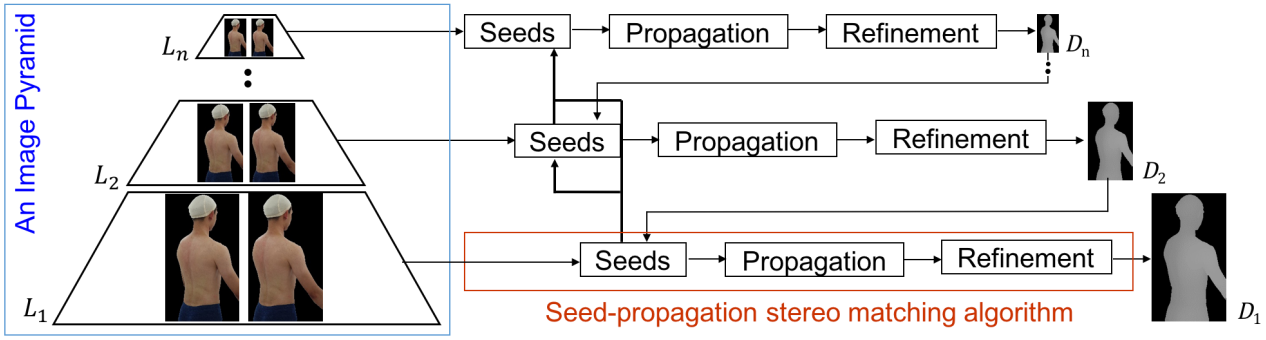


Figure 4: Flowchart of the proposed hierarchical seed-propagation stereo matching framework.

information including human skin features could be preserved as much as possible, benefiting depth recovery afterwards. The details of image pre-processing can be found in the supplementary material.

With an initial depth range of 1.5∼2.5 meters, a hierarchical framework is proposed to speed up the depth recovery. For each pair of stereo images, an image pyramid is built by down-sampling with a factor of two, as shown on the left of Fig. 4. $\{L_1, L_2, \ldots, L_n\}$ are used to indicate the layers, where n is the number of layers. As shown in Fig. 5, for the original input image with the resolution of 2880×3840, the lowest resolution layer L_n should be about 360×500 to preserve local salient features.

Figure 5: Point-clouds recovered from the hierarchical disparity images at 4 layers, starting from the coarsest layer L_4 (360×480) to the finest layer L_1 (2880×3840).

The seed-propagation stereo matching is conducted from top to bottom in the image pyramid (i.e., from L_n to L_1), which means the depth of pixels in the input images are estimated from coarse to fine (see Fig. 5). As shown in Fig. 4, to estimate a disparity image D_k for layer L_k, we apply a three-stage algorithm: matching seed extraction (Section 4.2.2), seed propagation (Section 4.2.3), and disparity image refinement (Section 4.2.4). First, multi-modal image information which is saved in multiple pairs of stereo images (as described in the supplementary material), are integrated to compute the matching cost volume V_k (for layer L_k). For each pair of stereo images corresponding to one type of photometric rendering parameters, a sub-cost-volume v_k is computed. V_k is computed by averaging all v_k for each pixel to gather comprehensive radiometric information. Matched features (i.e. matching seeds) are extracted based on V_k, and then used to guide the dense matching in low contrast regions.

In the proposed hierarchical framework, layers are closely related. First, to compensate the loss of information due to down-sampling operation, matchings seeds in higher resolution layers L_{k-1}~L_1 are down-scaled to the current layer L_k as a supplementary. When conflicts happen, the seeds derived from higher resolution layer are retained. Second, the disparity image D_{k+1} of the lower resolution layer L_{k+1} is used to reduce the searching volume for L_k, and speed up the disparity estimation. In addition, matches with extremely high confidence in L_{k+1} are scaled to L_k as candidates of the matching seeds. A pseudo algorithm of the proposed hierarchical stereo matching framework is presented as Algorithm 1.

4.2.2 Matching Seed Extraction

To tackle the lack of texture in human skin images, two steps are conducted in the proposed seed-propagation based stereo matching algorithm: (1) first extract robust matching features and (2) then let these robust matches guide the dense stereo matching. In this section, details of extracting matching seeds in one layer of the image pyramid are provided.

First, uniformly distributed features are extracted and matched. The input images are divided into 2D grids, and a certain number (i.e. 4 in the practical experiments) of blob and corner features are extracted in each patch using Difference-of-Gaussian (DoG) and Harris operator. The matching cost of Zero-Normalized Cross Correlation (ZNCC) $C(p_r, p_m)$ (definition in supplementary material) are used to choose the optimal matching pixels p_r and p_m, as it performs better for human skin images than other commonly used costs [19]. Based on the winner-take-all strategy in stereo matching, the best match p_m for a pixel p_r is selected by the largest value of $C(p_r, p_m)$.

However, due to the lack of texture of human skin, solely relying on ZNCC may not perform well. To ensure reliable matching seeds and correct dense matches, four constraints are added to determine if the best match $\{p_r, p_m\}$ can be accepted. *Photometric Consistency* which encourages reliable matching based on distinctiveness of a match from its neighboring matches. *Smoothness*

Algorithm 1 The hierarchical stereo matching framework

Input A pair of stereo images $\{I_{right}, I_{left}\}$;
Output A disparity image D $\{m_d\}$;
An image pyramid $P = \{L_1, L_2, ..., L_n\}$;
$i = n$
for all L_i in P **do**
  Computing a cost volume V_i for L_i;
  Extracting matching seeds S_i for L_i;
end for
Cost volumes $V = \{V_1, V_2, ..., V_n\}$
Matching seeds $S = \{S_1, S_2, ..., S_n\}$
for all L_i in P **do**
  if L_i is not L_n **then**
    Down-sampling matching seeds $\{S_{i-1}, ..., S_n\}$ to S_t
    $S_i \leftarrow S_i$ and S_t
  end if
  Computing dense stereo matching D_i for L_i
end for
return D^n

which ensures similar disparity between a pixel and its neighbors. *Ordering* which preserves the spatial relation between two neighboring matches. *Uniqueness* which guarantees the matching is commutative between reference image and matching image. Different from the stereo matching in [7], a metric "Confidence" is proposed to reserve the most distinct pixel matches which have been satisfied all the constraints. In addition, the importance of each constraint varies at different steps in the proposed seed-propagation based stereo matching. More details of the constraint definition are depicted in supplemental material. Fig. 6(a)∼(d) illustrate the extracted features. By checking four constraints sequentially, features at the same scanline in two stereo images are extracted and matched. Fig. 6(e) shows the matched features in red.

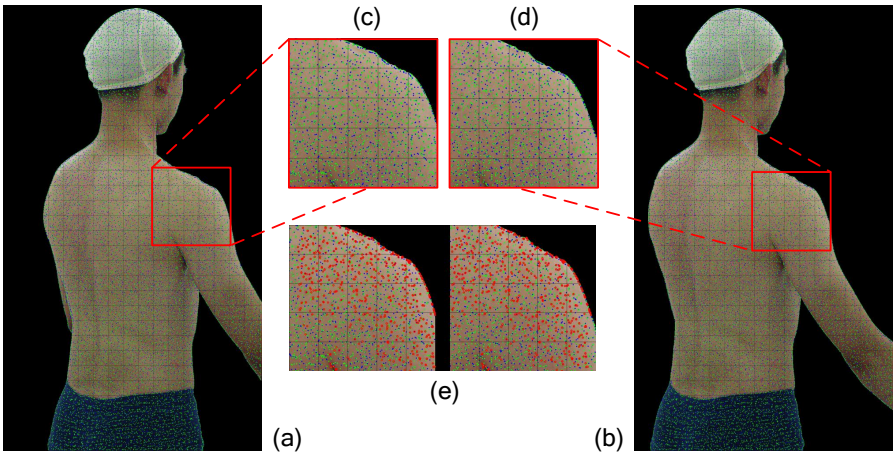


Figure 6: (a) and (b) show the blob (green) and the corner (blue) features extracted in the reference image and the matching image, respectively. To clearly demonstrate the uniform distribution of features, two corresponding cropped patches from two images are presented in (c) and (d). The matched features between them are shown in (e) with red color.

4.2.3 Seed-Propagation for Stereo Matching

The robust matched features, combined with feature matches derived from higher resolution layers (as presented in 1), are employed as matching seeds. Then, a best-first propagation strategy is performed to generate more matches in low contrast regions, starting from the neighboring areas of those seeds. A matching seed (p_r, p_m) is indicated as m_s, and a priority $P(m_s)$ is assigned as

$$P(m_s) = C(m_s) \cdot R(m_s) \tag{3}$$

which equals to the product of its matching cost and confidence.

The propagation starts from matching seeds with the highest priority. A new match generated in propagation also needs to satisfy the four constraints. Given textures are limited in most parts of human skin image, the smoothness constraint plays an important role in extracting accurate matches during propagation. The propagation terminates when no more candidate matches can be obtained, resulting in a quasi-dense disparity image, noted as D^Q.

After seed propagation, pixels fail to satisfy the above constraints remain un-matched. Based on the local smoothness of the human body shape, a disparity value $d(p)$ is assigned to an un-matched pixel p using image filtering as

$$d(p) = \sum_{q \in N(p)} W(q) * d(q) \tag{4}$$

where q is in the neighborhood of q, donated as $N(p)$. $W(q)$ is the filtering weight. The guided image filter, proposed in [18], is used to calculate the weight $W(q)$. Assume that the filtering output F and the guidance image G are linearly related (i.e.,

$F = a*G+b$), the guided image filter should ensure the consistency of gradient variation between F and G. By using the reference image I_r as the guidance image, the filtered disparity result D^F preserves features and edges in the reference image [18].

For now, an initial dense disparity result D^F of one layer is obtained. To further reduce the local matching ambiguities caused by the winner-take-all strategy, a dynamic programming based optimization [24] is employed to refine D^F along multiple directions to enforce global smoothness. In each direction r, the disparity $d(p)$ for pixel p in D^F is refined by minimizing the following objective function:

$$\sum_p C(p, p + d(p)) + \sum_p \lambda(p)\varphi(|d(p) - d(p + r)|), \tag{5}$$

which is the sum of the matching cost $C(p, p + d(p))$ and the penalization of disparity differences between the current pixel p and its adjacent pixels $p + r$ in direction r. λ is a weight function to control the degree of smoothness.

Fig. 7(a) shows 8 optimized directions in the dynamic programming. Since those constraints ensure the reliability of pixel matches, only $k(=16)$ disparity candidates near the initial $D^F(p)$ need to be computed during optimization. Instead of updating the disparity instantly after optimizing in each direction, the matching cost volume is accumulated for all directions to eliminate streak artifacts. Last, the final refined disparity D^R is obtained by applying general refinement techniques in stereo matching including region voting, cross-check and median filtering to remove outliers. A cross-based support areas $R(p)$ [31], as shown in Fig. 7(b), is used as the neighborhood. Besides, a sub-pixel enhancement technique which models disparity values and their matching costs as a quadratic polynomial function is used to compute the floating point disparity values.

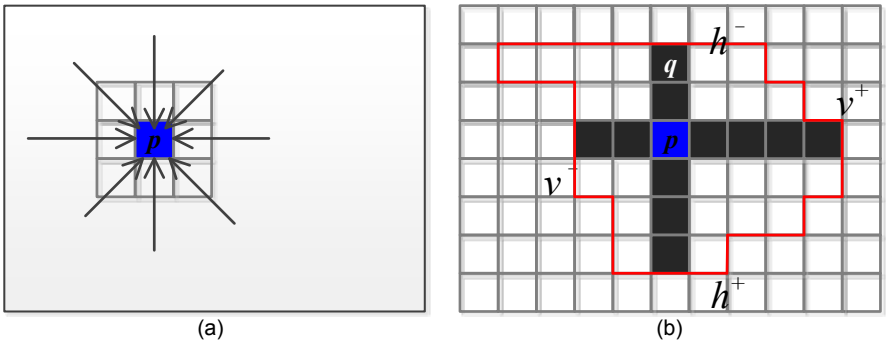


Figure 7: (a) Dynamic programming is performed in multiple directions to ensure semi-global smoothness. (b) Illustration of the cross based region R(p). For pixel p. h^-, h^+, v^- and v^+ are the left, right, top and bottom ranges, respectively.

Fig. 8(a) and (b) show the propagation result of the matching seeds extracted only in the current layer and the corresponding result supplemented with reliable matches from other hierarchical layers, respectively. The matches are obviously denser in the latter result. In Fig. 8, a RGB human body image is used as a base image to show the coverage of the resulting matches. Compared Fig. 8 (c) and (d), the depth discontinuity in (c) is improved a lot after the semi-global optimization.

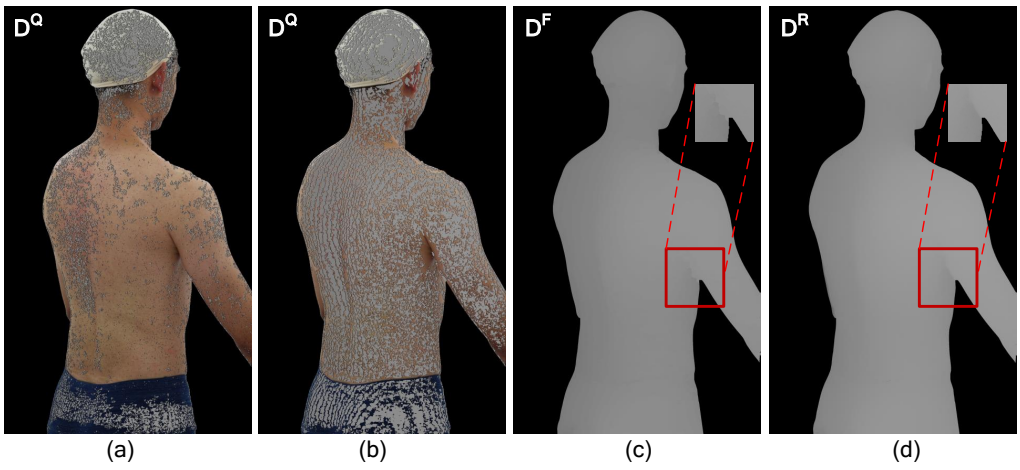


Figure 8: Disparity images generated in different steps by (a) propagating the matching seeds only extracted in the current layer; (b) propagating the matching seeds supplied with the reliable matches derived from other layers; (c) applying the guidance filter; and (d) semi-global dynamic programming optimization.

4.2.4 Human Body Disparity Refinement

Generally the shape of a human body can be modeled as a smooth parametric surface either for 2D disparity or 3D depth. Based on this prior, the obtained disparity image would be further refined in precision of float by applying two 3D geometry processing techniques tailored into 2D.

First, outliers with extremely large or small disparity could be rejected using statistical analysis [35] on the disparity image D^R. Without loss of generality, the mean and deviation of the difference between the pixel and its neighbors are assumed to be satisfied with the Gaussian distribution. A pixel with a mean difference value greater than a threshold will be rejected. In practice,

this progress iterates 3∼5 times, and those pixels with odd disparity values can all be removed. To illustrate the feasibility of the refinement, depth maps from disparity images are generated by using the triangulation formula:

$$Z = \frac{B * F}{D} \tag{6}$$

Here, B and F are the calibrated baseline and focal length, respectively. As shown in Fig. 9, outliers have been removed in the depth map after the refinement of statistical analysis.

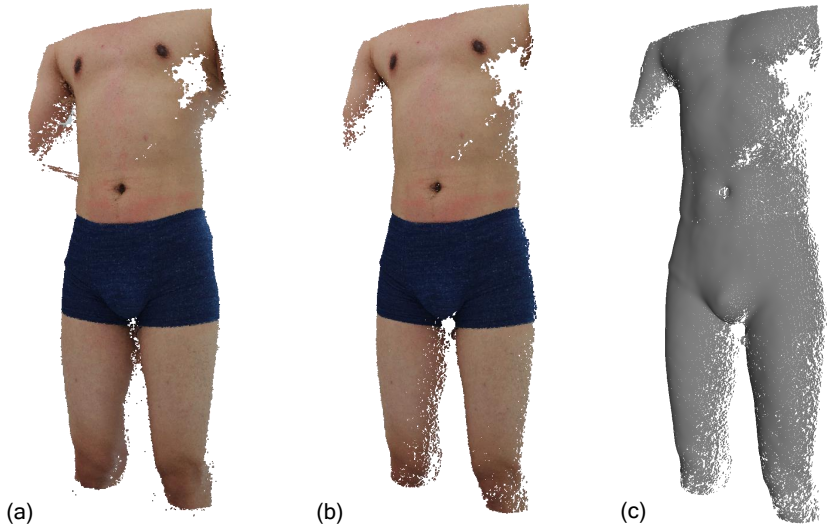


Figure 9: (a) and (b) are the recovered depth data before and after two refinement techniques: statistical outliers removal and adaptive moving least squares, respectively. (c) is (b) without textures.

After sub-pixel enhancement, matching errors would be magnified and lead to high frequency noises due to the inversely proportional relationship between depth and disparity. Adaptive moving least squares (AMLS) [2] technique is used in 2D disparity images to filter out the high frequency noises. The human body shape can be treated as a smooth parametric surface in 3D, meaning disparity values along a epipolar line (scanline) can be modeled as a smooth analytic function. For each scanline Y_c, pixels (x, y_c) are divided into several segments by constraining the disparity range of a segment within a threshold τ_d. Then a local polynomial $f(x, y_c)$ is used to model the relationship between the pixel coordinate (x, y_c) and its disparity $d(x, y_c)$. Finally, the disparity value $d(x, y_c)$ will be replaced by the fitted value of $f(x, y_c)$. τ_d is set to 5 in the practical experiments. As shown in Fig. 10, the fitted disparity curve is reasonably close to the integer disparity values. Moreover, as shown in Fig. 9(c), the depth map without texture is smooth and consistent with the human body geometry prior.

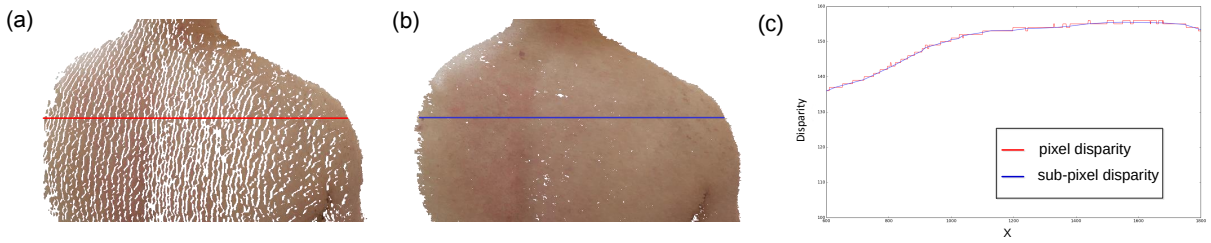


Figure 10: (a) and (b) are the recovered depth data from disparities before and after adaptive moving least squares. The discontinuities in (a) are greatly reduced. (c) Disparity values in a scan-line. The curves are with the same color as the two lines in (a) and (b).

4.3 Point Cloud Fusion and Surface Reconstruction

For each stereo rig in the proposed system, a clean and smooth partial point cloud can be generated from the disparity image by applying the triangulation formula and the refinement process (see Fig. 9(b)). To reconstruct a complete high-quality human body model, multi-view registration is employed to fuse partial point clouds under rigid and non-rigid transformations, and surface reconstruction is employed to fill missing data and generate watertight mesh model.

4.3.1 Multi-view Point Cloud Fusion

The partial data, denoted as $\{\mathcal{X}_1, \mathcal{X}_2 \ldots, \mathcal{X}_m\}$, needs to be fused into a complete human body point cloud $\mathcal{X}^H$. The fusion process contains three steps: an initial alignment based on global calibration in Section 4.1, a multi-view rigid registration, and a multi-view non-rigid registration.

First, $\mathcal{X}_i$ is transformed into a global coordinate system using the corresponding calibrated camera pose as $P_i \cdot \mathcal{X}_i$, resulting in the initial alignment result $\mathcal{X}^I$ which roughly forms the whole body (see Fig. 11(b)). For comparison, the coarse alignment result from bundle adjustment is shown in Fig. 11(a). It can be seen that the proposed global calibration largely improves the

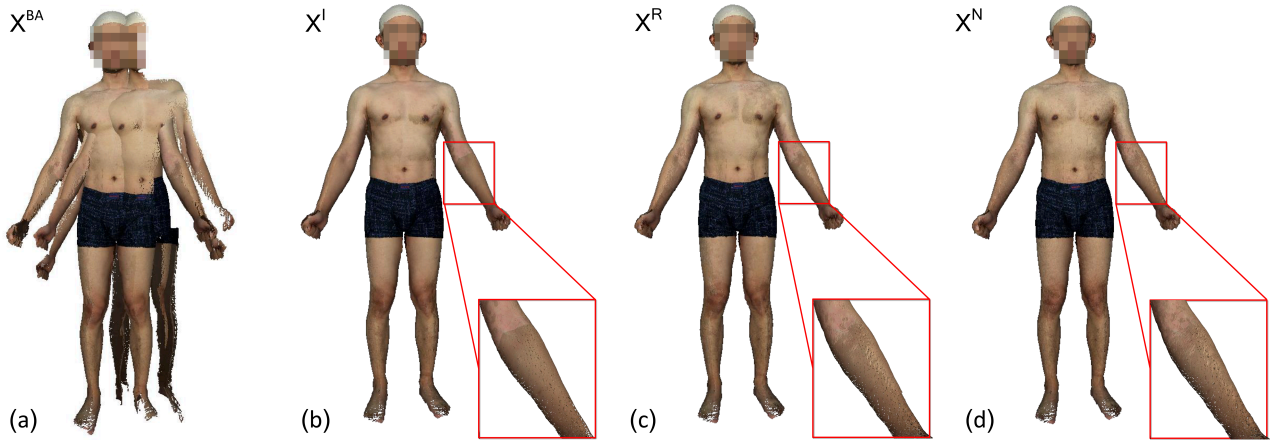


Figure 11: (a) The alignment based on the calibration result of the bundle adjustment. (b) The initial alignment based on our global camera calibration in Section 4.1. (c) and (d) show the better aligned point clouds after rigid and non-rigid registration, respectively.

initial alignment. Second, multi-view rigid registration is performed to minimize the distance between partial data, leading to an improved result $\mathcal{X}^R$ over initial alignment. The stereo rig setup of the proposed system is abstracted into an undirected spatial relation graph $\mathcal{G}$ with 30 nodes, each of which represents a partial point cloud from a stereo rig (see Fig. 12). Each graph edge connects two nodes with overlapping point clouds. Black edge connects two nodes with sufficient overlap, while for blue edge the overlapping condition should be checked. If the number of overlapping points are less than 10% of the whole point cloud, the corresponding edge should be removed.

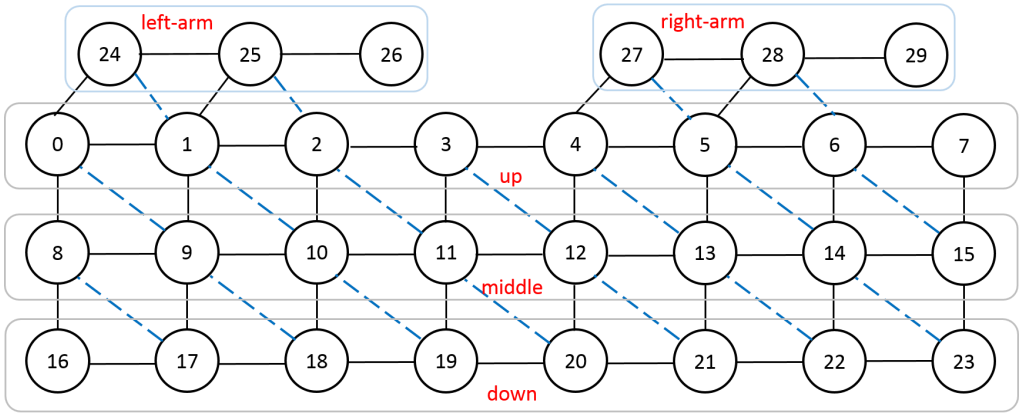


Figure 12: Illustration of the spatial relation graph $\mathcal{G}$.

Based on $\mathcal{G}$, a loop-based incremental registration algorithm in [45] is used to refine $\mathcal{X}^I$. The basic idea is to register all partial points loop by loop. Each loop corresponds to a spatial circular arrangement in $\mathcal{G}$. Once point clouds in each loop are merged, a process of global error diffusion is applied to distribute the residual error evenly to each point cloud. The improved result after rigid-registration is shown in Fig. 11(c), where points from left forearm are aligned more tightly.

Non-rigid registration is further employed to resolve geometry inconsistency caused by camera distortion and stereo matching deviation. As shown in Fig. 11(d), the points around the hands are cluttered because of non-rigid shape distortion of the same part. To tackle these artifacts, an improved hierarchical non-rigid registration method [11] is adopted to refine $\mathcal{X}^R$. The spatial relation in $\mathcal{G}$ is re-used and the warping functions are modeled as multiple thin plate splines. The final point cloud fusion result after non-rigid registration is denoted as $\mathcal{X}^N$, and shown in Fig. 11(d). Comparing the results in Fig. 11(c) and (d), it could be seen that the non-rigid registration result is much more compact, especially in boundaries of each point cloud. For more details of multi-view point cloud registration, please refer to [45, 47].

4.3.2 Surface Mesh Reconstruction

After data fusion, a complete point cloud of the human body shape is obtained. Due to self-occlusion, small holes caused by missing data may still appear in regions such as oxter, crotch, bottom of foot, and top of head. A hole filling step is introduced to fill the missing data based on template-based deformation [4]. In this research, it is tailored into a local approach since most parts of human body shape can be well captured. After filling all the missed data, de-noising and normal estimation are performed to polish the overlapping area, and prepare for the subsequent surface reconstruction. To generate a human body mesh with high fidelity, a multi-scale surface reconstruction method [46] is employed, where the reconstructed surface details are adaptive to the local curvatures of the captured point cloud. By properly deploying multi-scale B-spline basis functions on the adaptive signed distance field, the surface reconstruction problem can be reduced to a well-conditioned sparse linear system, which can be solved in a multi-grid way. Finally, a watertight human body mesh model can be generated. Detailed results will be demonstrated in the next section.

5 Evaluations and Discussion

The proposed system is evaluated on a number of subjects with different heights, weights and body shapes. Fig. 13 shows 13 resultant models, and the corresponding statistics are summarized in Table 1. In addition, Fig. 14 shows a reconstructed model in standard pose from three different views under different illumination conditions. It can be seen that global shape structure and local geometric details are reconstructed, and the normals for each point are smooth estimated. And several other poses of different subjects are shown in Fig. 15, which verify the feasibility of capturing non-standard T poses with varied self-occlusions. In the following, the precision of the captured models will be discussed and evaluated, as well as the precision of each component of the proposed system.

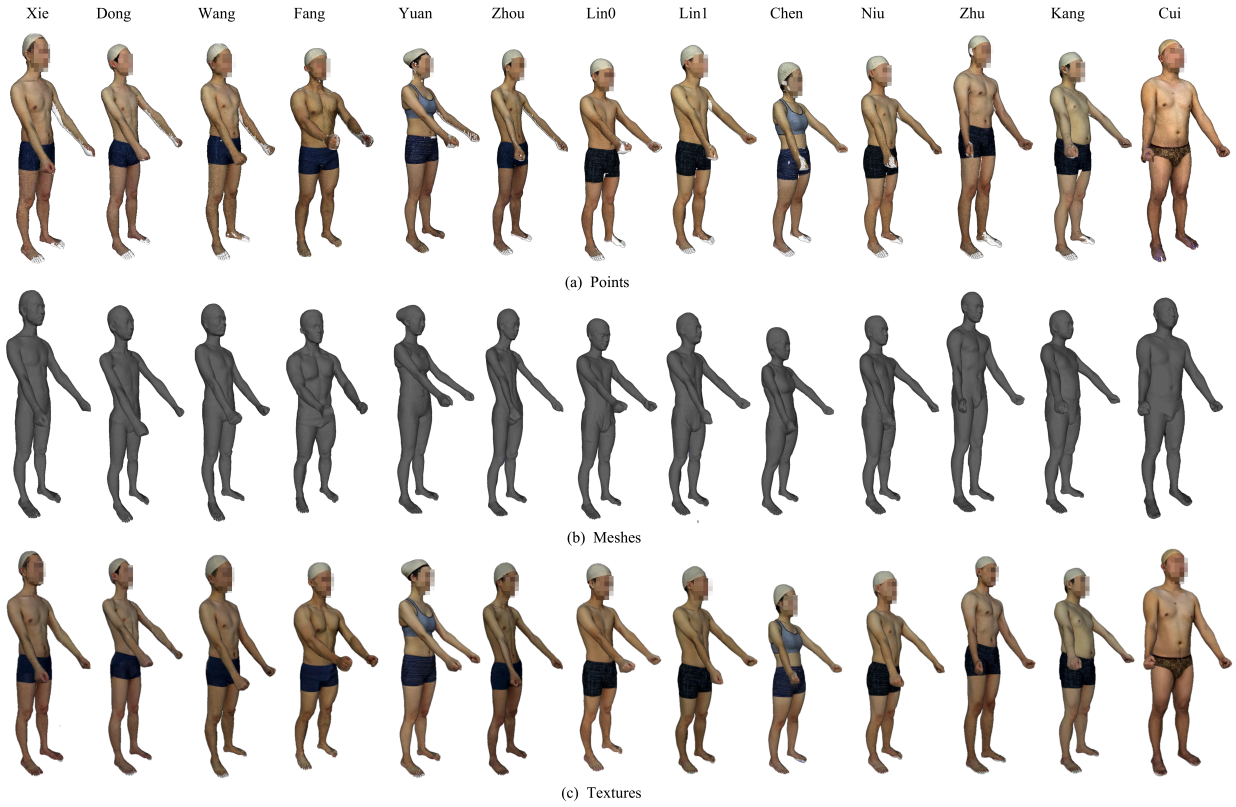


Figure 13: Results of the proposed system on 13 subjects. (a) the fused point clouds. (b) the reconstructed mesh geometry; and (c) the mesh model with texture information.

Name	G	H (cm)	W(kg)	#Points	#Triangles
Xie	M	178	75	4,931,672	7,216,784
Dong	M	172	53	5,723,490	7,115,008
Wang	M	176	72	5,873,831	7,918,656
Fang	M	173	61	5,893,414	7,754,780
Yuan	F	170	50	5,754,611	7,193,872
Zhou	M	173	62	5,614,031	7,234,976
Lin0	M	171	66	5,785,377	7,519,738
Lin1	M	174	68	5,737,954	7,444,508
Chen	F	161	46	5,675,333	7,497,878
Niu	M	168	58	5,720,679	7,493,470
Zhu	M	189	78	5,984,364	7,524,286
Kang	M	174	77	5,118,980	7,002,914
Cui	M	183	81	5,922,790	7,143,504

Table 1: The statistics of captured models. G (gender), H (height), W (weight), #Points (fused points), #Triangles (model faces).

5.1 Accuracy of the Single Point-cloud

The local calibration in Section 4.1 and the stereo matching in Section 4.2 are two main factors that influencing the accuracy of partial point cloud from a single stereo rig.

First, the rectification error [10] is employed to evaluate the accuracy of local calibration for stereo rigs. Based on row-aligned epipolar geometry, given a match (p_1, p_2), the rectification error ϵ_r is defined as $\epsilon_r = ||v(p_1) - v(p_2)||$, where $v(p_1)$ and $v(p_2)$ are

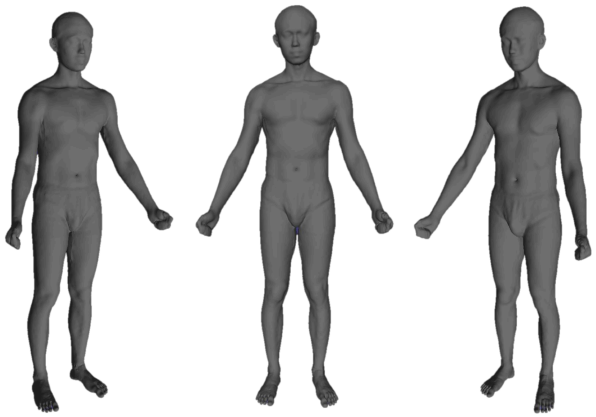


Figure 14: Example of a reconstructed model under different illumination conditions and viewpoints.

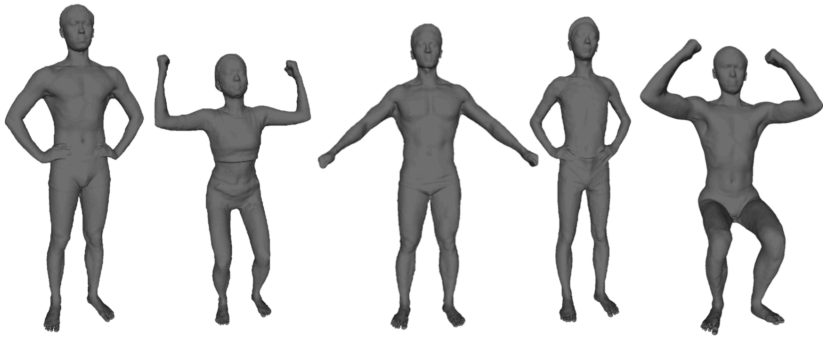


Figure 15: Example of different poses.

the row coordinates of p_1 and p_2. In practice, ϵ_r is evaluated at the corners of the checkerboard pattern for each stereo rig. The average of ϵ_r equals to 0.12 and the maximum value is 1.14 at the pixel level, which is about $0.006mm$ to $0.01mm$ at the distance level. The largest error usually occurs at the image border without covering human body. Therefore, the row-aligned epipolar geometry is guaranteed for later binocular stereo matching.

Second, the accuracy of recovered depth from stereo matching can be estimated using Eq. (7), based on the triangulation rule in epipolar geometry. It evaluates the depth error due to the mis-match of one pixel:

$$\Delta Z = \mathcal{B} \cdot \mathcal{F} \cdot max\{\frac{1}{d} - \frac{1}{d \pm 1}\}, \tag{7}$$

where $\mathcal{B}$ is the baseline length and F is the focal length. This metric varies with different depth values. In the proposed system, the captured human usually stands within a depth range of $1.5{\sim}2.5m$. $\mathcal{B}$ is set to $150mm$. And $\mathcal{F}$ is approximately $50.2mm$, which is estimated by the local calibration in Section 4.1. Thus, the error of recovered depth ranges from 1 to 3 millimeters according to Eq. (7), meaning that one pixel mis-match in disparity results in an average error of $2mm$ in depth, which lays a foundation for the precision of the whole system.

5.2 Accuracy of the Acquired Human Body Models

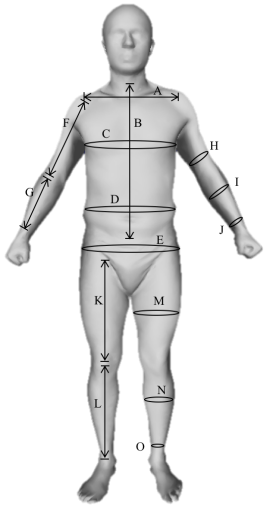
In this paper, several commonly-used anthropometry measurements, as shown in Fig. 16, are utilized to evaluate the accuracy of acquired human body models.

Among these anthropometry measurements, body lengths and body circumferences represent the reconstruction error of a single point cloud and the reconstruction error of the complete registered point cloud, respectively. The precision of a single point cloud could be regarded as the accuracy of the proposed stereo matching algorithm. And, the precision of the complete registered point is influenced by the initial alignment from global calibration (Section 4.1) and the following registration of point clouds (Section 4.3).

It is challenging to assess the ground-truth values of these measurements during the capture process, considering the non-rigidity of the human body. This article tests the method of obtaining these values by manual measurement. Testing results show that the manual measurement is neither reliable nor consistent. During experiments, 10 participants are recruited to measure the anthropometry parameters of a plastic mannequin. The average measurement variation is $6.23mm$. In addition, one participant is asked to measure the same human body for ten times. The average measurement variation is $4.75mm$. Notably, the manual measurement can be either subjective, or easily affected by human body variation due to unwanted movement.

To resolve the above issues, this research proposes a novel evaluation method that obtains simultaneous anthropometry measurements while acquiring human body models. In the proposed method, thin sticky measuring tapes are attached to human body to obtain ground-truth anthropometry measurements (listed in Fig. 16). And the human body with tapes is acquired and reconstructed by the proposed system. Two subjects (one male, one female) are used as representatives for evaluation, and the male representative is shown in Fig. 17.

Then, the reconstructed anthropometry values are calculated as the Euclidean distances between a set of points along the tape contours, shown as in Fig. 18. In the meantime, the ground-truth of corresponding measurements can be directly read



	Measurements
A	Shoulder breadth
B	Shoulder/crotch length
C	Chest circumference
D	Waist circumference
E	Hip circumference
F	Upper arm length
G	Fore arm length
H	Upper arm circumference
I	Fore arm circumference
J	Wrist circumference
K	Thigh length
L	Calf length
M	Thigh circumference
N	Calf circumference
O	Ankle circumference

Figure 16: Illustration of anthropometry measurements.

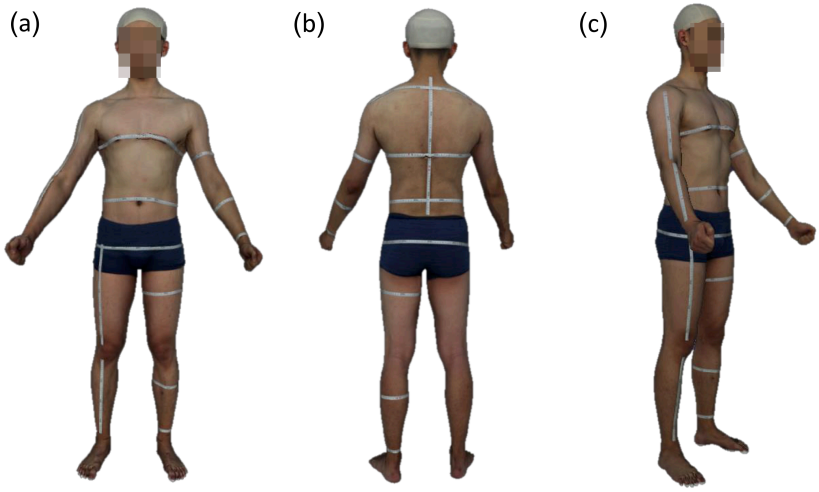


Figure 17: (a), (b) and (c) show the acquired human body model of a male subject with attached measuring tapes from the front, back and side view, respectively. The measuring tapes are well captured.

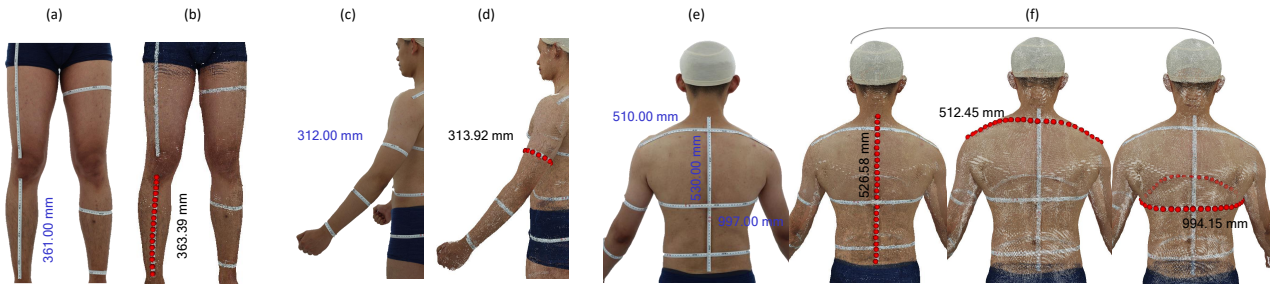


Figure 18: (a) and (b) show the captured image and the calculated length of the right calf, respectively. (c) and (d) show the same for the left upper arm circumference. (e) and (f) show the image and the calculated lengths of the shoulder/crotch, shoulder breadth and chest circumference (from left to right), respectively.

from 2D images. This simultaneous measurement and acquisition approach allows accurate evaluation of captured anthropometry parameters. Taking the measurement of right calf length for example (see Fig. 18(a)), the ground-truth value of the right calf length is 361.00mm (numbers from the raw image). The reconstructed value is 363.39mm by computing the Euclidean distances along the tape on the reconstructed model (see Fig. 18(b)). Hence the reconstruction error is 2.39mm.

	Female				Male			
M	IGC	Ours	PS	COL	IGC	Ours	PS	COL
A	6.71	2.94	10.61	6.82	7.05	2.45	10.14	5.23
B	4.46	1.24	10.38	8.91	5.61	3.22	8.35	4.02
C	9.75	**5.04**	10.33	5.55	4.90	2.65	9.73	4.43
D	5.13	3.24	11.21	10.12	6.34	**4.54**	10.84	9.87
E	7.82	2.19	14.35	9.35	5.12	2.35	9.49	6.14
F	6.19	1.82	9.19	8.17	4.68	2.24	8.64	9.31
G	4.93	1.71	7.42	9.46	7.81	2.15	9.11	6.48
H	6.61	2.25	8.16	11.31	2.32	1.92	8.4	4.20
I	3.28	2.34	9.78	7.92	6.95	2.02	8.13	9.01
J	8.11	2.47	13.32	10.65	7.87	1.48	9.57	5.78
K	11.09	2.89	12.35	4.37	6.56	1.65	8.83	3.25
L	7.53	2.29	9.67	8.65	5.14	2.39	9.48	9.12
M	8.53	3.12	7.87	5.78	8.29	2.16	10.09	7.62
N	3.84	1.41	10.18	9.54	10.14	3.08	12.69	10.51
O	5.01	2.28	8.15	6.49	9.07	2.18	10.95	8.26

Table 2: Statistics of reconstruction errors. IGC, PS and COL are short for the initial global calibration, Photoscan and COLMAP, respectively.

Table 2 shows the statistics on the reconstruction accuracy for all measurements. The error is the difference between the ground truth value read from the color image and the measured values from the reconstructed human model. The unit is millimeter. The average error for our results is 2.457mm, while the max error is up to 5.04mm occurred for the chest circumference. The reason is that the underarm parts have missing data and are reconstructed by hole filling. Except the human face, there are few textures in the human body surface. From Fig. 16 and Fig. 17, it can be seen almost all of the anthropometry measurements locating in the texture-less body regions. The measurement A,B,C,D,E,K and M in human torso, which locate from the shoulder to the knee, are collected separately to clarify the effectiveness in the reconstruction of low texture areas. As listed in Table 2, the max reconstruction error in low texture areas is 5.04mm for the female and 4.54mm for the male, which is the same as the statistics of all the measurements. And the average texture-less reconstruction error is 2.95mm for the female and 2.71mm for the male, respectively, which is larger than all the measurements. In addition, we evaluate the reconstruction error of the initial alignment from global calibration, of which average error is 6.56mm. It is consistent with the visual comparison of the alignment result in Fig. 11 that the reconstruction error is decreased after registration. Also, based on the statistics, we find that slimmer human body covering less capturing space leads to bigger reconstruction error. This is due to the fact that larger depth range results in more accurate stereo matching.

5.3 Comparison

With measuring tapes attached on two representatives, this research compares the proposed method with the state-of-art commercial software called Photoscan [1], of which kernel is derived from PMVS [14], a representative general-purpose reconstruction method based on multi-view geometry. With additional engineering optimizations, Photoscan can produce much better reconstruction than the initial PMVS. The highest accurate level is set for each step in the work-flow of the Photoscan. Due to insufficient textures of human skins, inaccurate estimation of surface patches has been created such as non-smoothness and outliers during the estimation of dense point-clouds by the multi-view stereopsis.

The proposed method is also compared with COLMAP [39], which is one of the best multi-view stereo (MVS) pipeline for general objects models acquisition. It takes the output of structure-from-motion (SfM) to compute depth and normal information for every pixel in 2D image, then utilizes the depth and normal maps of multiple images to produce a dense point cloud of the scene/object. Based on the fused point cloud, screened poisson surface reconstruction [22] is adopted to reconstruct the surface geometry. It should be noticed that the SfM step in COLMAP is failed due to insufficient features of the human skin. Only a subset of images is resolved (half of all the images at most), which leads to incomplete human body shapes. By feeding global calibration results obtained by the proposed method into the COLMAP, the reconstructed human body models are generated, as shown in Fig. 19(c). However, due to incorrect normal estimation during MVS optimization, the reconstructed models are severely contaminated by noises.

Fig. 19 shows the qualitative comparison of three methods. It could be seen that the geometry information recovered by the proposed method is much more similar to the real human body surface than other two methods. The quantitative comparison on the anthropometry measurements is performed to the reconstructed models generated by Photoscan and COLMAP, respectively. The comparison of statistics can be found in Table 2, which demonstrates the advantage of the proposed work. Similarly, the average texure-less reconstruction error of Photoscan is 11.01mm for the female and 9.63mm for the male, respectively, while COLMAP is 7.27mm for the female and 5.79mm for the male, respectively. By comparison, the proposed method improves the reconstruction accuracy in the low texture areas of human body surface.

5.4 Evaluation on the Benchmark

Following the standard way of evaluating a passive multi-view stereopsis system, the proposed system is evaluated on two benchmark datasets "temple" and "dino", which are provided by the Middlebury Multi View Stereo [36]. There are 312 images and 363 images in "temple" and "dino", respectively. In order to integrate the datasets into the proposed multi-binocular pipeline, two adjacent images are grouped into a pair of stereo images and rectified with the given calibration parameters. Taking the "temple" dataset for example, 166 pairs of stereo images are organized to generate dense depth information via the proposed

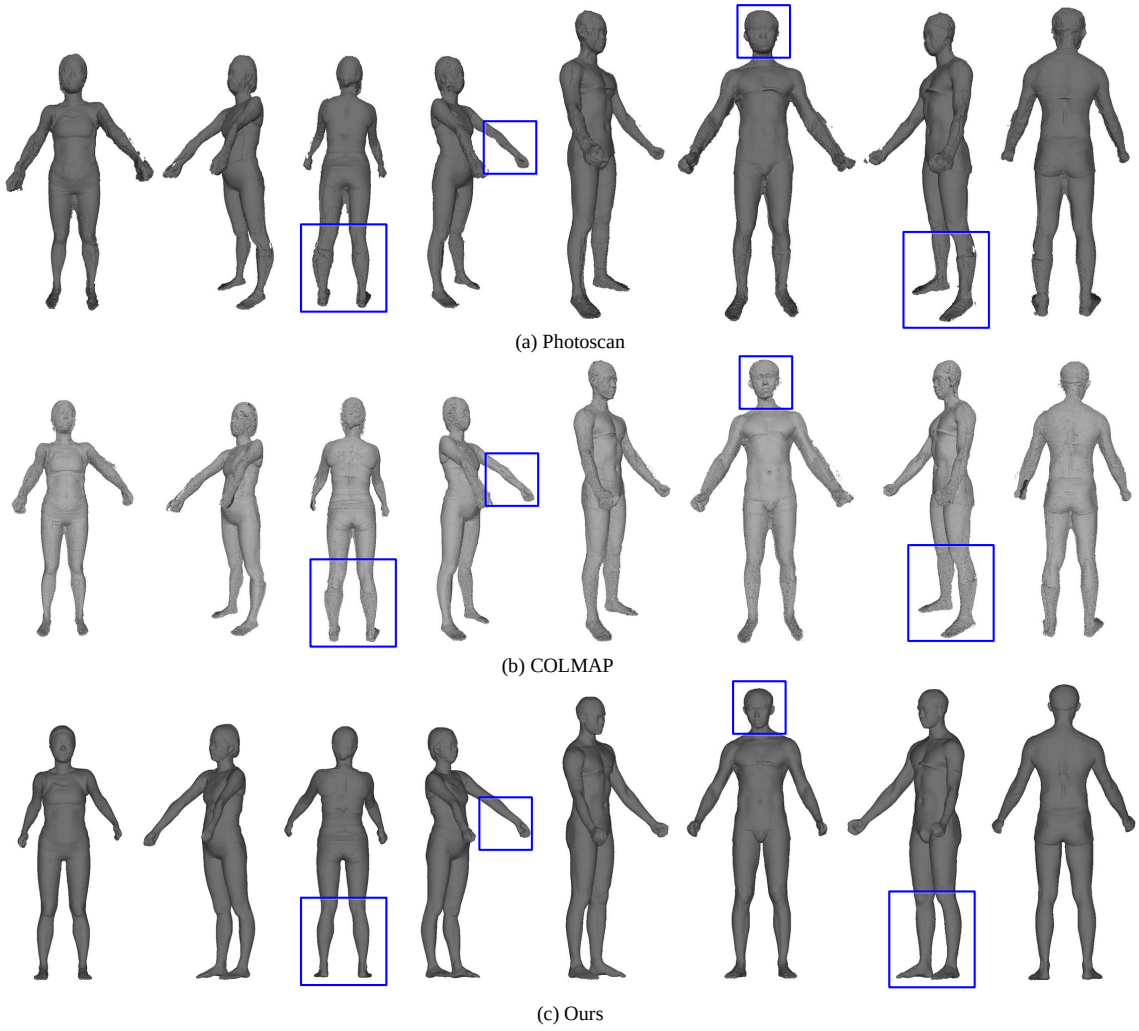


Figure 19: Reconstructed models obtained (a) by Photoscan. (b) by COLMAP. (c) by the proposed method.

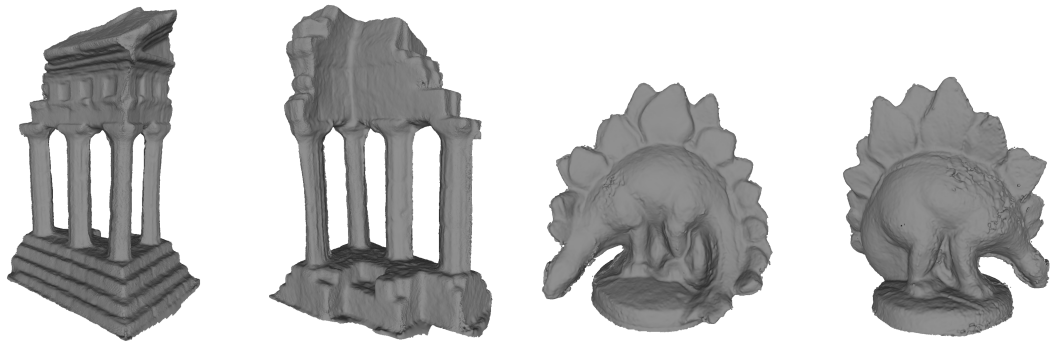


Figure 20: Evaluation of the proposed system on the Middlebury Multi-View Stereo dataset. (a) acquired model of "Temple". (b) acquired model of "Dino".

stereo matching algorithm. Finally, all partial depth information is fused via the proposed multi-view point clouds registration. The topology graph used in the fusion is obtained by calculating overlaps between point clouds. As shown in Fig. 20, detailed and completed static models for two datasets could be acquired by the proposed system.

5.5 Performance

The entire acquisition pipeline is executed on a desktop PC with 3GHz CPU and 16GB memory. The average computational time of each step is shown in Table 3. Overall it takes about 30 minutes to generate a mesh model for a static human body. The most time-consuming part is multi-view registration, because the high density of partial point clouds produced by stereo matching and the registration algorithm is iterative. Using a subset of points could improve the computational efficiency but may sacrifice result

step	time (min.)
global calibration	1
seed-growing stereo matching	1.5
refinements of disparity registration	3
multi-view point clouds registration	15
hole filling	5
3D surface reconstruction	5.5

Table 3: The execution time of each step in the pipeline.

quality. The stereo matching algorithm is conducted on 30 stereo rigs parallelly to generate point clouds since the stereo rigs are independent of each other, so as the disparity image refinement step.

5.6 Ablation Studies

Two ablation studies are discussed in this section, including the influence of camera numbers and the robustness of the proposed seed propagation.

5.6.1 Numbers of Cameras

An incremental experiment is conducted to study the influence of the number of cameras. During the experiment, there are 5 tests with increasing number of viewpoints. Based on the hardware configuration of the proposed system, there are two selection rules in each test. First, a pair of stereo cameras, which is regarded as the depth sensor, is the smallest unit to be selected. Second, 6 pairs of stereo cameras located in diagonal viewpoints (as shown in Fig. 1) will be added incrementally in these tests. Namely, there are 12, 24, 36, 48 and 60 cameras in 5 tests respectively. The acquired completed point clouds in the incremental experiment are shown in Fig. 21.

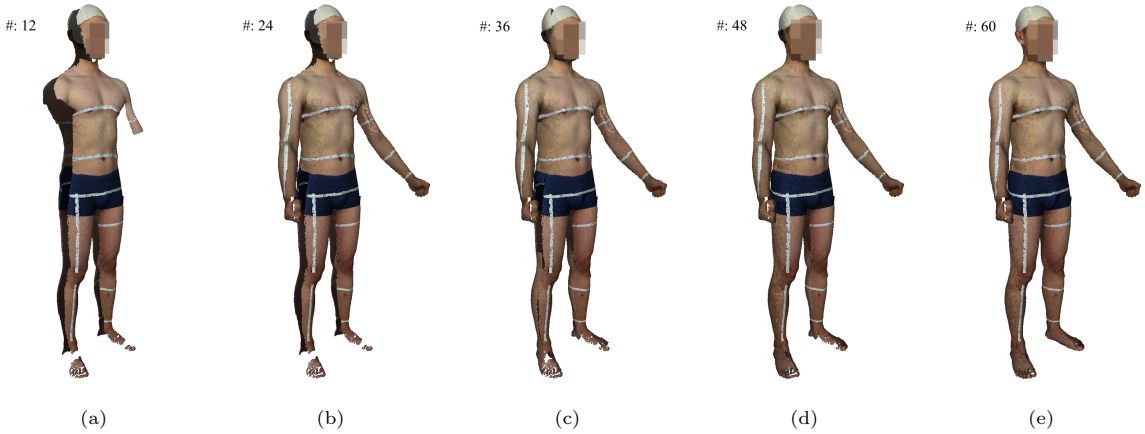


Figure 21: Completed point cloud acquired, respectively, by (a) 12 cameras. (b) 24 cameras. (c) 36 cameras. (d) 48 cameras. (e) 60 cameras.

To avoid large distortion caused by wide baseline, the baseline of each stereo rig is set to $18cm$. Due to the relatively small baseline, capturing data could be missed in some parts of the human body if fewer cameras are used, which also can be observed in Fig. 21. From Fig. 21 (a) to (e), with the addition of cameras focusing on different areas of the human body, the resulting point cloud is achieved more completely. As a conclusion, the amount of cameras in the proposed system is sufficient to capture a complete human body robustly, expect for a few self-occluded areas. Moreover, more cameras could provide more comprehensive texture information (see the measuring tapes in Fig 21), which will benefit the accuracy measurements of reconstructed human body models.

5.6.2 Seed Propagation

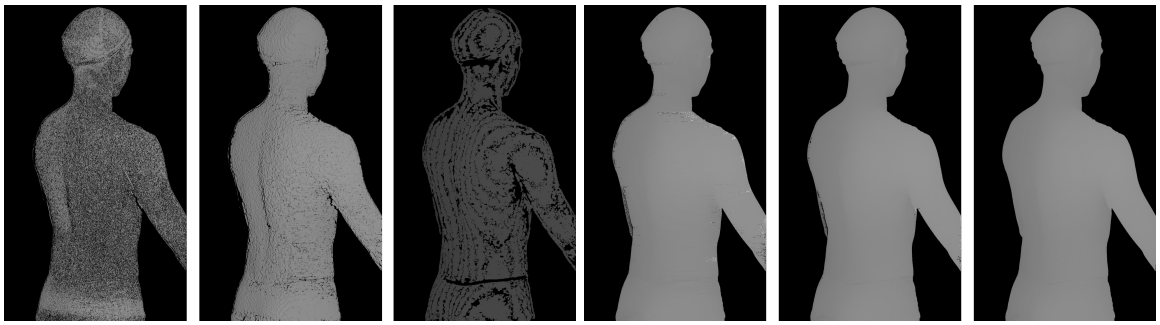
In this section, robustness and effectiveness of the proposed seed propagation based stereo matching method are discussed. At first, the input stereo images are divided into 2D grids and then 4 features are extracted in each grid, see Fig. 6. In this case, sufficient and uniformly distributed salient pixels are extracted to be matched as matching seeds. The number of matching seeds depends on values of two thresholds in the photometric consistency constraint, namely, τ_c for the ZNCC matching cost $C(\cdot)$ and τ_r for the match reliability $R(\cdot)$ in Eq. 3. The larger the two threshold values, the less the number of matching seeds. During experiments, to ensure the reliability of matching seeds, τ_c is set to 0.95 (of which maximum is 1.0) and τ_r is set to 1.5 (a match could be considered very credible with the value greater than 1.1). To clarify the influence of the number of matching seeds, three different threshold values of τ_c and τ_r are tested. Besides the above empirical threshold values, other two test sets are $\{\tau_c = 0.5, \tau_r = 0.85\}$ and $\{\tau_c = 0.99, \tau_r = 1.9\}$. Results of matching seeds of three different sets are shown in Fig. 22(a). Quasi-dense disparity images before subsequent refinement process are shown in Fig. 22(b) From left to right, the matching seeds are shown in gray-scale pixels which corresponding to increasing threshold values, respectively. It can be concluded that smaller τ_c andτ_r should generate a number of matching seeds but also introduce more mismatches.

After extraction of matching seeds, seed propagation is conducted to generate quasi-dense disparity results. The whole process follows the rule of priority prorogation, which is, neighboring matching seeds with higher credibility $P(\cdot)$ (see Fig. 3) are matched in advance. Considering the lack of texture in those neighboring areas of matching seeds, in practical, the photometric consistency

(a) (b)

Figure 22: (a) matching seeds with $\{\tau_c = 0.5, \tau_r = 0.85\}$, $\{\tau_c = 0.95, \tau_r = 1.5\}$ and $\{\tau_c = 0.99, \tau_r = 1.9\}$, respectively. From left to right, numbers of matching seeds are 3468, 1980 and 646, respectively. (b) the corresponding quasi-dense disparity results generated by seed propagation.

constraint is relaxed to $\{\tau_c = 0.75, \tau_r = 1.05\}$. Similar to the ablation study on matching seeds, three sets of thresholds with increasing values are tested. Besides the practical values, other two sets are $\{\tau_c = 0.35, \tau_r = 0.85\}$ and $\{\tau_c = 0.95, \tau_r = 1.5\}$,. It should be noted that the latter set is the same as the values in the extraction of matching seeds, which is very strict. Three quasi-dense disparity images are shown in Fig. 23(a) and the corresponding final disparity images are shown in Fig. 23(b). With too strict thresholds in the propagation, most pixels remains unmatched in the quasi-dense result. After the refinement process, final disparity results of three different threshold sets are well obtained. And, too loose or too strict propagation thresholds increase the time cost of refinement process. To balance off the algorithm performance and the time cost, $\tau_c = 0.75$, $\tau_r = 1.05$ has been used in the seed propagation process of all the practical experiments. Eventually, as shown in Fig. 13(a), point clouds are completely recovered by the proposed method for various input capture images.



(a) (b)

Figure 23: (a) quasi-dense disparity results with $\{\tau_c = 0.35, \tau_r = 0.85\}$, $\{\tau_c = 0.75, \tau_r = 1.05\}$ and $\{\tau_c = 0.95, \tau_r = 1.5\}$, respectively. From left to right, numbers of matching seeds are 3468, 1980 and 646, respectively. (b) the corresponding final disparity results after the refinement process.

6 Conclusion

A multi-view high-precision human body acquisition system is proposed in this work. The average reconstruction accuracy is within $2.5mm$ in terms of anthropometry measurements. The hardware setup is based on consumed DSLR cameras which avoid the significant cost of high-end scanning devices and the interference between commodity RGBD sensors with limited accuracy. The acquisition pipeline is subsequently designed as following: (1) first calibrate both local stereo rigs and global camera array; (2) then recover dense and precise point clouds via novel hierarchical stereo matching; (3)finally reconstruct high-quality watertight surface mesh model. This research tests the proposed system by capturing a number of human body models with varied heights, weights and body shapes. A novel evaluation method is proposed that prevents the measurement errors when producing ground-truth anthropometry parameters. The results and comparisons clearly justify the performance of our system over the state-of-the-art approaches.

Limitations The proposed system is designed to capture human body shapes with precise anthropometry parameters, but not a specific part as human face/hand. However, the proposed acquisition strategy would be useful therein as well with tailored camera setup either separates from or builds upon the present system. Also, the local-global calibration is dedicated to the proposed system and lays the foundation for accurate depth recovery, but more complicated than typical MVS-based system without binocular stereo rigs.

Future Works In the future, the joint optimization between binocular and multi-view stereos will be further explored to recover more accurate point clouds. More efficient geometry processing and surface reconstruction methods for dense point clouds will also be investigated [25].

7 Acknowledgements

This work was supported jointly by the National Natural Science Foundation of China under Grants No. 61732015 and No. 61472349, Key Research and Development Program of Zhejiang Province under Grant No. 2018C01090.

References

[1] AgisoftLLC. Agisoft photoscan user manual: professional edition, 2018.

[2] M. Alexa, J. Behr, D. Cohen-Or, S. Fleishman, D. Levin, and C. T. Silva. Computing and rendering point set surfaces. *IEEE Transactions on visualization and computer graphics*, 9(1):3–15, 2003.

[3] T. Alldieck, M. A. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. *arXiv preprint arXiv:1803.04758*, 2018.

[4] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: reconstruction and parameterization from range scans. In *ACM transactions on graphics (TOG)*, volume 22, pages 587–594. ACM, 2003.

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

[6] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[7] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross. High-quality single-shot capture of facial geometry. In *ACM Transactions on Graphics (ToG)*, volume 29, page 40. ACM, 2010.

[8] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2300–2308, 2015.

[9] D. Bradley, T. Boubekeur, and W. Heidrich. Accurate multi-view reconstruction using robust binocular stereo and surface meshing. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[10] D. Bradley and W. Heidrich. Binocular camera calibration using rectification error. In *Computer and Robot Vision (CRV), 2010 Canadian Conference on*, pages 183–190. IEEE, 2010.

[11] B. J. Brown and S. Rusinkiewicz. Global non-rigid alignment of 3-d scans. In *ACM Transactions on Graphics (TOG)*, volume 26, page 21. ACM, 2007.

[12] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (ToG)*, 34(4):69, 2015.

[13] A. Feng, E. S. Rosenberg, and A. Shapiro. Just-in-time, viable, 3-d avatars from scans. *Computer Animation and Virtual Worlds*, 28(3-4):e1769, 2017.

[14] Y. Furukawa and J. Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376, 2010.

[15] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1381–1388. IEEE, 2009.

[16] M. Habbecke and L. Kobbelt. A surface-growing approach to multi-view stereo reconstruction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[17] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. In *Computer Graphics Forum*, volume 28, pages 337–346. Wiley Online Library, 2009.

[18] K. He, J. Sun, and X. Tang. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6):1397–1409, 2013.

[19] H. Hirschmuller and D. Scharstein. Evaluation of cost functions for stereo matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[20] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2307–2315, 2017.

[21] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015.

[22] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG)*, 32(3):29, 2013.

[23] H. Kim and A. Hilton. 3d scene reconstruction from multiple spherical stereo pairs. *International journal of computer vision*, 104(1):94–116, 2013.

[24] J. C. Kim, K. M. Lee, B. T. Choi, and S. U. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 1075–1082. IEEE, 2005.

[25] P. Koppel. Agisoft photoscan: Point cloud accuracy in close range configuration, 2015.

[26] V. Leroy, J.-S. Franco, and E. Boyer. Multi-view dynamic shape refinement using local temporal integration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3094–3103, 2017.

[27] V. Leroy, J.-S. Franco, and E. Boyer. Shape reconstruction using volume sweeping and learned photoconsistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 781–796, 2018.

[28] M. Lhuillier and L. Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE transactions on pattern analysis and machine intelligence*, 27(3):418–433, 2005.

[29] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev. 3d self-portraits. *ACM Transactions on Graphics (TOG)*, 32(6):187, 2013.

[30] Y. Liu, Q. Dai, and W. Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE transactions on visualization and computer graphics*, 16(3):407–418, 2009.

[31] X. Mei, X. Sun, M. Zhou, S. Jiao, H. Wang, and X. Zhang. On building an accurate stereo matching system on graphics hardware. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 467–474. IEEE, 2011.

[32] J. J. Moré. The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer, 1978.

[33] L. Quan. *Image-based modeling*. Springer Science & Business Media, 2010.

[34] F. Remondino. 3-d reconstruction of static human body shape from image sequence. *Computer Vision and Image Understanding*, 93(1):65–85, 2004.

[35] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz. Towards 3d point cloud based object maps for household environments. *Robotics and Autonomous Systems*, 56(11):927–941, 2008.

[36] D. Scharstein and R. Szeliski. Middlebury multi view stereo page, 2006.

[37] D. Scharstein and R. Szeliski. Middlebury stereo evaluation-version 2. *The Middlebury Computer Vision Pages (online), available from (accessed 2015-03-02)*, 2011.

[38] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[39] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[40] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer vision and pattern recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528. IEEE, 2006.

[41] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *IEEE international conference on computer vision*, pages 915–922, 2003.

[42] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE computer graphics and applications*, 27(3):21–31, 2007.

[43] R. Szeliski. A multi-view approach to motion and stereo. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 1, pages 157–163. IEEE, 1999.

[44] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[45] Y. Tang and J. Feng. Hierarchical multiview rigid registration. In *Computer Graphics Forum*, volume 34, pages 77–87. Wiley Online Library, 2015.

[46] Y. Tang and J. Feng. Multi-scale surface reconstruction based on a curvature-adaptive signed distance field. *Computers & Graphics*, 70:28–38, 2018.

[47] Y. Tang, S. Luo, Q. Ran, J. Kang, and J. Feng. Multi-view non-rigid registration based on multiple thin-plate splines. *Journal of Computer-Aided Design & Computer Graphics*, 29:2153–2161, 2017.

[48] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2009.

[49] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan. Scanning 3d full human bodies using kinects. *IEEE transactions on visualization and computer graphics*, 18(4):643–650, 2012.

[50] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustmenta modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[51] T. Tung, S. Nobuhara, and T. Matsuyama. Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1709–1716. IEEE, 2009.

[52] C. Wu, B. Wilburn, Y. Matsushita, and C. Theobalt. High-quality shape from multi-view stereo and shading under general illumination. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 969–976. IEEE, 2011.

[53] M. Ylimäki, J. Kannala, J. Holappa, S. S. Brandt, and J. Heikkilä. Fast and accurate multi-view reconstruction by multi-stage prioritised matching. *IET Computer Vision*, 9(4):576–587, 2015.

[54] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 910–919, 2017.

[55] C. Zhang, S. Pujades, M. J. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5484–5493, 2017.

[56] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.