Video-Based Outdoor Human Reconstruction

Hao Zhu, Yebin Liu, Jingtao Fan, Qionghai Dai, *Senior Member, IEEE*, and Xun Cao, *Member, IEEE*

***Abstract*— A human body scanning system of great practical convenience, which can be used in an outdoor environment, is proposed. The system uses only a single conventional video camera without the aid of special sensors or controlled illuminations. We leverage the structure from motion calibration results directly and improve the available video-based dense 3D reconstruction by integrating the surface smoothness constraints. The point cloud reinforcement is proposed to detect and adjust the conflict point data for the slender and shaky body parts. Combined with the silhouette adaptation, the proposed point cloud reinforcement achieves reasonable and plausible mesh reconstruction on these challenging parts. We further introduce the close-shot frames to refine the prereconstructed mesh model, leading to a colored watertight model. The overall system is approximate to automatic since only one or two times of painting brush interaction are required for robust and high-quality multiview image segmentation. The experiment results on various test sequences demonstrate the effectiveness and the robustness of the proposed method, even under very challenging scenarios when shaking body, varying illumination, and textureless regions occur.**

***Index Terms*— 3D reconstruction, close shot, surface reinforce, video-based, watertight model.**

I. Introduction

IN RECENT years, there are growing demands for the 3D model of human bodies in various fields, ranging from movie/TV industry, electronic games, and virtual reality to 3D printing. Human body scanning has many applications in computer vision and industry standardization. The traditional 3D scanning techniques, e.g., laser scanner, require sophisticated equipments, which are expensive, sensitive to calibration error, time-consuming, and hard to operate. The consumer level 3D sensors, e.g., Kinect, are increasingly popular, since they provide a more convenient approach for amateurs to reconstruct the 3D models. However, the specific hardware is still required, and the depth quality greatly degrades for outdoor scenarios.

Stereo-based 3D reconstruction from a video or an image sequence has been the subject of intense research for its flexibility. In general, the stereo-based methods fall into two main categories, i.e., active and passive. The active methods, such as laser scanner [1], structured illumination [2], and even method that interacts with the target objects [3], usually provide higher reconstruction accuracy, but may fail in outdoor environments with strong sun light. Instead, the passive methods recover the 3D model from several regular photographs, which presents more practical benefit for the outdoor 3D reconstruction. The structure from motion (SFM) algorithm has been studied well for the past decades, and can be applied to a variety of realistic scenes. Several image-based reconstruction applications or software, including Visual SFM [4] + PMVS [5]/CMP-MVS [6], PhotoScan,[1] and 123D Catch,[2] have achieved 3D model generation from photos, and do well in static object reconstruction. However, as the illumination of the outdoor scenes is uncertain and changeable, 3D reconstruction in outdoor environment tends to be affected by the illumination change and limited image quality. Furthermore, the human body contains many delicate, slender, and shaky parts, like legs, feet, arms, and hands, which make the human reconstruction outside the room even more fragile. Therefore, recovering the 3D model of a live human in outdoor environment with a portable camera still remains a very challenging problem, especially when the target human cannot maintain absolutely static during the video capture.

In this paper, a stereo-based method without requiring any special devices or capture skills is proposed, as shown in Fig. 1. The method reconstructs the 3D model of humans from a short piece of video or an image sequence taken around the target person. Inspired by the monocular dense tracking method [7], we extract matching information from the input video/images to overcome the instability of the outdoor scenes. A robust 3D model estimation is proposed to reconstruct the human model from the video taken with ubiquitous outside environment. The proposed method consists of an off-line processing that only uses a short piece of video as input. This kind of video clips can be obtained with much flexibility by using a digital video (DV) camera or even a small cell phone. The video clip can be captured by a walking person or a moving vehicle around the target object, which only takes dozens of seconds. A simple interaction step is introduced to segment the human body accurately. The overall processing merely needs one or two interactions, so the system works in an approximately automatic manner. The reconstructed 3D models can also be printed using a 3D printer for potential practical applications. In sum, we make the following technical contributions.

Manuscript received September 8, 2015; revised February 20, 2016; accepted July 22, 2016. Date of publication July 28, 2016; date of current version April 3, 2017. This work was supported in part by the National Key Foundation for Exploring Scientific Instrument under Grant 2013YQ140517, in part by the National NSF of China under Grant 61522111, Grant 61371166, Grant 61422107, and Grant 61531014, and in part by the NSF of Jiangsu Province, China under Grant BK20130583. This paper was recommended by Associate Editor P. Eisert. *(Corresponding authors: Yebin Liu; Xun Cao.)*

H. Zhu and X. Cao are with the Department of Electrical Science and Technology, Nanjing University, Nanjing 210023, China (e-mail: caoxun@nju.edu.cn).

Y. Liu, J. Fan, and Q. Dai are with the Automation Department, Tsinghua University, Beijing 100084, China (e-mail: liuyebin@mail.tsinghua.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2016.2596118

[1] http://www.agisoft.com

[2] http://www.123dapp.com/catch

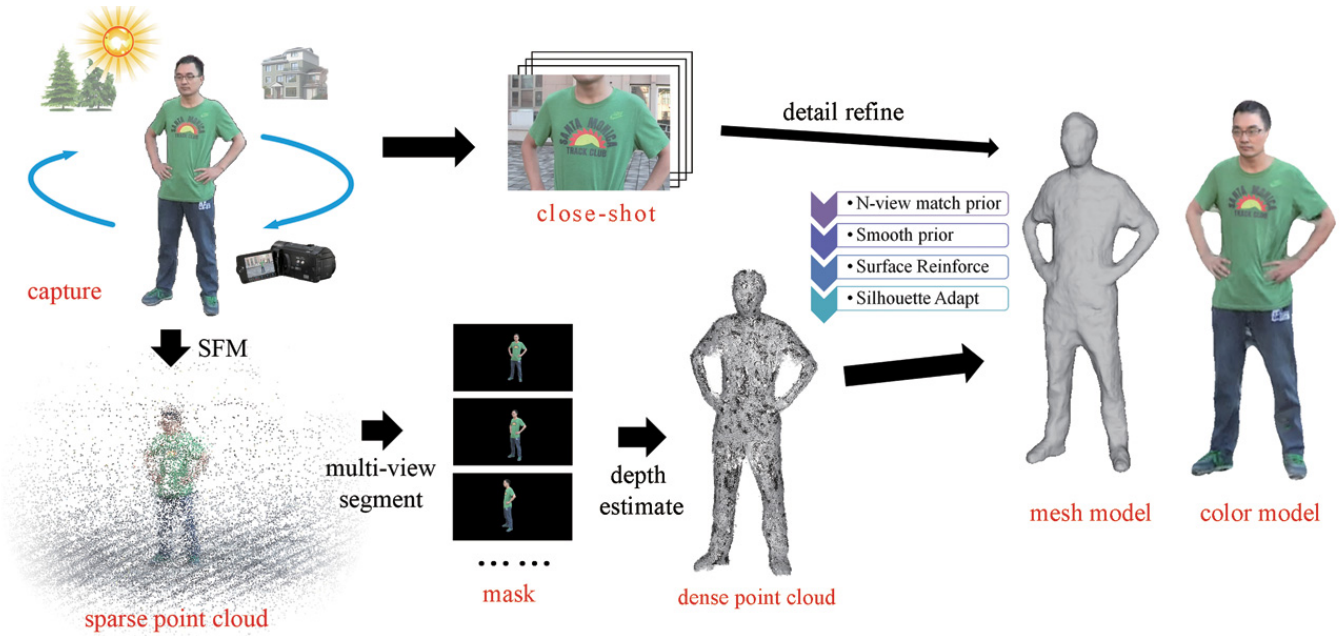


Fig. 1. Whole pipeline of our system. The overall system is approximate to automatic, as only one or two times of painting brush interaction are required to generate a high-quality human model.

1) We directly leverage the SFM camera calibration results and improve the available video-based dense 3D point cloud reconstruction by integrating the surface smoothness constraints.
2) To improve the accuracy on the shaky body parts, we propose the point cloud reinforcement to detect and adjust the conflict point data for the shaky body parts. Combined with the silhouette adaptation, the proposed point cloud reinforcement achieves reasonable and plausible mesh reconstruction on these challenge parts.
3) To further improve the surface details, we propose the close-shot refinement including the automatic grouping and mesh deformation, driven by reliable pixel as well as a detail enhancement optimization based on the close-shot video.

II. Related Work

The proposed scheme is dedicated to outdoor human 3D reconstruction, which is closely related to various literature aiming at multiview stereo (MVS) and 3D modeling. We will review the previous work in these categories.

A. Multiview Stereo

We could refer to the benchmarks in [8] and [9] for a comprehensive survey. Furukawa and Ponce [5] put forward the photo consistence-based optimization method, and considered the 3D points as a patch. Liu *et al.* [10] integrated silhouette information and epipolar constraint into the variational framework for continuous depth map estimation. Li *et al.* [11] introduced a bundle optimization method and used a DAISY feature to compute depth map. The algorithm performs well on the outdoor data sets when the camera parameters are given. The traditional MVS algorithms advance in high reconstruction precision, but most of them are not suitable for complex illumination and outdoor scene.

B. Structure From Motion

Pollefeys *et al.* [12] put the work on reconstruction with a handheld camera, then in [13], a real-time video-based 3D acquisition system was presented. The system collects video streams, as well as GPS and inertia measurements to reconstruct the model of urban scenes. Newcombe *et al.* [7] proposed the method for real-time interactive 3D reconstruction, which relies on dense, pixelwise methods. Their system applies the global optimization to reconstruct a fine surface of the scene. Stühmer *et al.* [14] presented a similar work estimating the depth maps from multiple views and converting them to triangle meshes based on the respective neighborhood connectivity. Both research are facing a plane scene and cannot generate a close model. Kolev *et al.* [15] proposed an efficient and accurate scheme for the integration of multiple stereo-based depth measurements, allowing to obtain the 3D models of pleasing quality interactively and entirely on device. Although their work realized the real-time reconstruction, the results are relatively discrete point clouds, while this paper aimed at reconstructing an integrated, closed human mesh model.

C. Prior

As prior knowledge is significant in image-based 3D reconstruction field, we discuss the prior knowledge separately. Furukawa and Ponce [16] presented to construct a coarse surface approximation in the form of a visual hull. Both geometric constraints and photo consistency constraints are enforced to acquire 3D shapes. Gall *et al.* [17] presented to use silhouette as prior. After the skeleton is built, the approximate surface skinning, true small scale deformations, or nonrigid garment motion are captured by fitting the surface to the silhouettes. Cremers and Kolev [18] tackled the reconstruction problem as the one of minimizing a convex function where the

exact silhouette consistency is imposed as a convex constraint, which restricts the domain of the admissible functions. Bao *et al.* [19] incorporated semantic information in the form of learned category-level shape priors and object detection.

D. Passive Human Modeling

The image-based body modeling methods have drawn much attention in the past decades. Template aided methods including [17], [20], and [21] employed multiview video to guide the prescanned model or articulated mesh deforming. These methods are suitable to motion tracking and rendering, but the accuracy of recovered shape is undefined. Wu *et al.* [22], [23] took advantage of photometric stereo to recover detailed geometry. However, the methods have the strong assumption on a reflectance model, relying on controlled illumination or uniform surface appearance, which could only be achieved in the studio. Low-cost body scanning systems, including [24]–[26], were designed using multiple fixed camera system to reconstruct a human body. These systems achieve high accuracy and show the feasibility of close range photogrammetry in medical application. More about the body scanners for anthropometric data are reviewed in [27].

E. Range Sensor

As range sensor provides credible depth information, the main study focuses on dense points alignment, fusion, and deformation. Newcombe *et al.* [28] presented a Kinect fusion system for accurate real-time mapping of the complex and arbitrary indoor scenes in variable lighting conditions. Amateurs are able to generate the 3D models of target objects with stunning details and accuracy. Tong *et al.* [29] presented a scanning system for capturing the 3D full human body models using multiple Kinects. Three parts of the body are scanned separately and registered under nonrigid deformation. Chang and Zwicker [30] introduced the pairwise nonrigid registration techniques to handle different types of deformations such as quasi-articulated motions, and in [31], they used the global alignment method to cope with larger deformations. Li *et al.* [32] developed an automatic pipeline that allows nonexpertise users to capture complete and fully textured 3D models of themselves in minutes, using only a single Kinect sensor. Barmpoutis [33] proposed to reconstruct the 3D model of the human body from a sequence of Red Green Blue-Depth frames. The framework runs in real time and allows the human subject to move arbitrarily in front of the camera. Comparing with an active scanning method, our method requires only one video recorder, which is more available than the depth sensors. The low-cost range sensors like Kinect and Xtion are in relatively low space resolution than an image sensor, and they tend to fail in outdoor scenes due to the sunlight interference.

Our method differs from all kinds of aforementioned literatures, relaxes all the assumptions or control on illumination conditions, and neither requires any special sensors or add-on equipments. Besides the flexibility, by taking advantage of both multiview segmentation and the reinforcement of the object surfaces, the proposed method is able to generate a watertight 3D model robustly in spite of the object shaking or illumination varying.

III. Point Cloud Reconstruction

In this section, we will illustrate each step in the pipeline of the depth map estimation. We first capture a video of an outdoor human using a handheld camera (DV or cell phone). The capture process takes ∼20–30 s to move around the object human. If higher 3D reconstruction accuracy is expected, another dozens of seconds are consumed to shoot at the details in a closer distance. We then use an SFM method to calibrate the images and segment the several keyframes using the N-view match points. The depth images of the keyframes are estimated using a modified cost volume [7], and several priors are employed to refine them.

A. Calibration

In the first stage of the modeling pipeline, the video is decoded into image sequence, and all the frames in the sequence are calibrated using SFM. As introduced in Section II, an SFM method has been studied in the past decades, and standard pipeline has been presented, such as Visual SFM tools [4], [34] and OpenMVG.[3] Our system adopts OpenMVG to perform the SFM algorithm. Meanwhile, we have verified the feasibility of visual SFM result to support our system. The calibration phase accomplishes the following pipelines: first, the SIFT [35] features of the image sequence are extracted and matched; then, an incremental SFM [36] is adopted to generate the calibration parameters. The calibrating result is refined using bundle adjustment, and the points that could be observed in more than three views are stored as the N-view matches. Given the camera parameters, a sparse point cloud of the scene is recovered from the N-view matches. The density of the sparse cloud depends on the quantity of the SIFT features in the image. More specifically, there are more dense points in a textured region, and the scattered points in a repeated or ambiguous texture region. After that, we do not intend to reconstruct the model from the sparse cloud, but use the points as a constraint in the depth image estimation.

B. Segmentation

After calibration, several keyframes are extracted from the image sequence. The keyframes are chosen at regular intervals to ensure appropriate baseline length. The interval frame number is 15 in all our experiments, so that we could obtain 20–30 keyframes in an object-centered circle, and the segmentation is applied merely on these keyframes.

The keyframes are segmented using the multiview segmentation approach that is similar to [37]. As demonstrated in Fig. 2, users are allowed to interact with the image segmentation program, indicating the foreground and background with the paint brush. The selecting part of image transports the label to the point in N-view point cloud. The foreground probability is propagated from one point to another according to the connectivity and the nearest neighbor method. Then, the N-view point cloud is projected to each keyframe; meanwhile, the foreground probability is transferred from the point to the pixels of other keyframe images. Graph cut method [38] is

[3] http://openmvg.readthedocs.org/en/latest/openMVG/openMVG

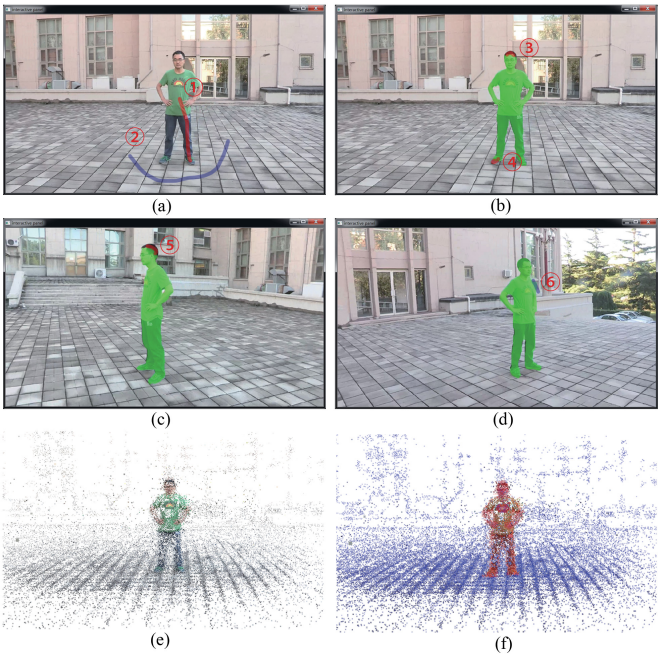


Fig. 2. Segmentation process includes an interaction with users. In (a), the user picks out the foreground with red paintbrush and background with blue paintbrush. In order to make a preferable initial segmentation, it is better to select the object human foot as foreground and the floor near the feet as background. The probability is propagated to another view via m-view point cloud. A rough segmentation result in every keyframe is returned to the user, and the optional revision is then supplemented by the user. The keyframes are shown in sequence, if the segmentation of current image is ideal, this one is passed, and if the segmentation has a problem, like (b)–(d), the user should make a revision as (3)–(6). Here, the green area is the segmentation result in previous iteration. In general, one or two revisions on head or limbs are enough to make a fine segmentation. In (e), the N-view point cloud is generated by Visual SFM. The point cloud is too sparse and noisy, but is enough to guide the multiview segmentation. (f) Labeled point cloud. Red: foreground points. Blue: background points.

adopted to segment the image according to the probability map. The users are allowed to make a revision on images during the iteration of the segmentation process. In general, one or two revisions, which focus on object's head and arm are enough to make a fine segmentation result.

In our experiments, the segment results are generally fine, while few tiny-range misjudgments occur on the edge of the mask. The impact of misjudgments on subsequent processing is weak, because the segmentation results are used mainly in two phases.

1) In depth estimation phase, we assign the depth computing region according to the segmentation. The exceeding misjudgment pixels will lead to monstrous depth, which can be easily filtered in the fusion phase. The filtered depth vacancy is replenished by the depth image of the other views.
2) In the silhouette adaption phase, as the displacement of surface is controlled by the multiview silhouette in iteration, the tiny-range errors of segmentation are dispersed in the experiment.

C. Dense Depth Estimation

Our stereo matching process is a modification of the cost volume method in [7]. We gather the contextual frames of one keyframe to compose a cluster $\mathcal{M}$, which shares a plenty of overlapping regions. A projective photometric cost volume is defined for each keyframe in Fig. 3, and we allocate a list of depth candidates to each pixel in the mask, and each candidate contains a matching cost of its depth. Then, the matching costs are computed by projecting the point of the current pixel and depth to the other images in the bundle, and summing the L_1 norm of the photometric errors. Mathematically, the match cost is formulated as

$$C(u,v,d)=\frac{1}{m}\sum_{i\in\mathcal{M}}\| I_{\text{key}}(u,v)-I_i(P_i(invP_{\text{key}}(u,v,d)))\|_1 \tag{1}$$

where $invP_i(u,v,d)$ reconstructs the point in the world axis from the pixel location(x,y) and depth d, and $P_i(x,y,z)$ projects the point to the image coordinate. Both the functions adopt the calibration parameters of frame i. m is the number of frames in the cluster $\mathcal{M}$, and key is the index of the keyframe. $I_i(u,v)$ is the sum of L_1 norm of R, G, and B intensity difference of the pixel (u,v). After every depth candidate's cost of one pixel are calculated, a winner-take-all strategy is adopted to select the depth with the lowest cost. The difference between our approach and the cost volume in [7] is that we use the N-view match point cloud and segmentation as prior to set float search origins for different pixels in one image, and the depth searching range is determined by the scale of the m-view point cloud and the image size, which trades off the efficiency and precision. In our experiment, the depth interval is the N-view match point cloud height divided by the pixel number of mask image in the vertical direction.

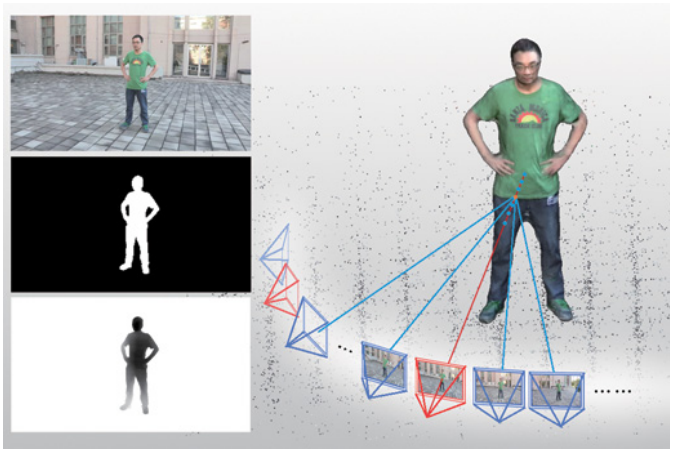


Fig. 3. Cost volume model of the depth map estimation. From the top–down of the left part: one keyframe image (red frames), segment result (mask), and initial depth map. The points on the red line crossing the keyframe's camera center denote the depth candidates. Blue rays: projection relationship between the candidate points and the pixels in the reference frame (blue frames).

D. Depth Refinement

The depth map of the previous phase is rough and scattered in the textureless regions, and thus, we employ several priors to constrain the cost model.

1) Shape Prior: As the shape information is obtained (N-view match point cloud and keyframe segment), we apply

a flexible range of cost volume for each pixel. To be specific, the first step is to estimate the initial depth using all the keyframes' masks. A ray crossing the camera center and the pixel in the world coordinate is drawn. The points on the ray are projected to the other keyframes from the origin point (camera center). If the point is the first, which falls into all other keyframes' masks, we set its depth as initial depth. Next, the N-view point clouds are applied to replace the initial depth. As the N-view points are unevenly distributed and contain a mass of outliers, we elect the average depth of the dense part to alter the initial depth in the first step.

2) Smoothness: Smoothness is the most common prior used in the image-based 3D reconstruction. As for human body, we intend to improve the surface consistency with the guidance of photometric consistency. Based on the disparity map refinement in [39], the depth map refinement iteration is applied to smooth the depth map and improve the subscale accuracy. The depth value of the pixel d is updated in every iteration

$$d' = (\omega_p d_p + \omega_s d_s)/(\omega_p + \omega_s) \tag{2}$$

where ω_p represents the match consistency

$$\omega_p = \begin{cases} c_0 - c_{-1} & (c_{-1} < c_0, c_{+1}) \\ 0.5(c_{-1} + c_{+1} - 2c_0) & (c_0 < c_{-1}, c_{+1}) \\ c_0 - c_{+1} & (c_0 < c_{-1}, c_{+1}) \end{cases} \tag{3}$$

where c_0 is the match cost of the selected depth, c_{-1} and c_{+1} are the match cost of the depth next to the selected depth, forward and backward separately. d_s is a smooth term and d_p is a linear fitting value of the matching cost, with

$$ds = \frac{\omega_x(d_{x-1,y} + d_{x+1,y}) + \omega_y(d_{x,y-1} + d_{x,y+1})}{2(\omega_x + \omega_y)} \tag{4}$$

$$d_p = \frac{(d_1 - d_0)\|c_1 - c_0\|_1 + (d_{-1} - d_0)\|c_{-1} - c_0\|_1}{\|c_1 - c_0|_0 + \|c_{-1} - c_0\|_1} + d_0. \tag{5}$$

3) Connectivity: Naturally, each part of the human body is interlinked, and it is impossible to find more than one isolated closed mesh in a single human model. However, a few depth map-based 3D reconstruction algorithms perform badly in thin area of human body like arms and legs because of the difficulty to identify the outliers at the slender part. On the other hand, a few outliers may make the constructed mesh (e.g., through Poisson reconstruction [40]) break down. Another labile factor is that the human object could hardly maintain a pose during the video recording. A little waggle would disturb the depth estimation and lead to the break in the thin part. We will demonstrate how we ensure the connectivity of human model and solve the problem of body shaking in Sections IV-A and IV-B.

IV. Surface Reinforce and Refinement

In this section, we explain how to reinforce the model to ensure the integrity and how to tackle the shaking problem. First, the depth maps are fused, and the outliers are removed by the approach in [41]. Then, the point clouds reinforce is applied to adjust the point in the conflict regions and avoid the crack in slender parts. A mesh model is generated using *Poisson Reconstruction* [40], and we adopt silhouette adaption to refine the mesh model. Finally, we are able to improve the model details with the close-shot clips of the captured video, and render the entire model with Poisson blending algorithm [42].

A. Point Cloud Reinforcement

After we obtained several keyframe depth maps in Section IV, the point clouds are generated by inverse-projecting the point of each pixel into the world coordinate. The points that keep away from the points of contextual keyframe are judged to be outliers and are removed. According to the experiment, most of the outliers near the flat regions are removed, but in curvature regions like arm, only the outliers outside the cylindroid are detected, and the bad matched points inside the cylindroid are remained, because they could find the supporting points inside the dense curvature structure. Moreover, the bare arms and legs are slender parts with few textures, where the bad matched points are inclined to be gathered. Fig. 4(b) and (c) demonstrates the fact that the shaking region produces the crack of legs as well.

We propose to adjust the point with the prior that the structure of a human body is connected and reinforces a cylinder-like surface. The main idea is to 'push' points inside the cylinder-like structure out in iteration. The points of the keyframe are corresponding to each pixel of the segmented masks, so the point clouds contain the reliable silhouette information to the corresponding keyframe. Here is the detailed procedure.

First, we pick out the curved regions with conflicts. The principal component analysis algorithm is used to compute the point's normals in a certain point clouds. We call the point clouds generated from a certain frame as target clouds and the all others' clouds as reference clouds. A K-D tree is built to fast search the points of reference clouds near one point of the target clouds. Focusing on a certain point of the target clouds, if the intersection angle of its normal and a neighbor point's normal is greater than 90°, i.e., the normalized cross correlation of the two normals is less than zero, this neighbor point conflicts with the specified point, and the specified point is judged to be a conflict point when the conflict ratio in the neighborhood surpasses a threshold κ. In all our experiments, $\kappa = 0.3$, and Fig. 4(a) (left) shows the checked conflict points.

After picking out the conflict points, we apply the iteration only on the conflict region. The ith vector-type point in keyframe n is $\Upsilon_n(i)$, and the normal of the point is $\eta_n(i)$. The iteration is given by

$$\Upsilon'_n(i) = \frac{\sum_{j \in N(i)} \Upsilon'_n(j) + \alpha \sum_{h \in N(i)} \Upsilon_m(h)\lambda(i,h)\mu(i,h)}{N_j + N_h} \cdot \eta_n(i) \tag{6}$$

$$\lambda(i,h) = \begin{cases} 1 & \text{if } (\Upsilon_m(h) - \Upsilon_n(i)) \cdot \eta_n(i) > 0 \\ 0 & \text{else} \end{cases} \tag{7}$$

$$\mu(i,h) = \begin{cases} 1 & \text{if } \eta_n(i) \cdot \eta_n(h) > 0 \\ 0 & \text{else} \end{cases} \tag{8}$$

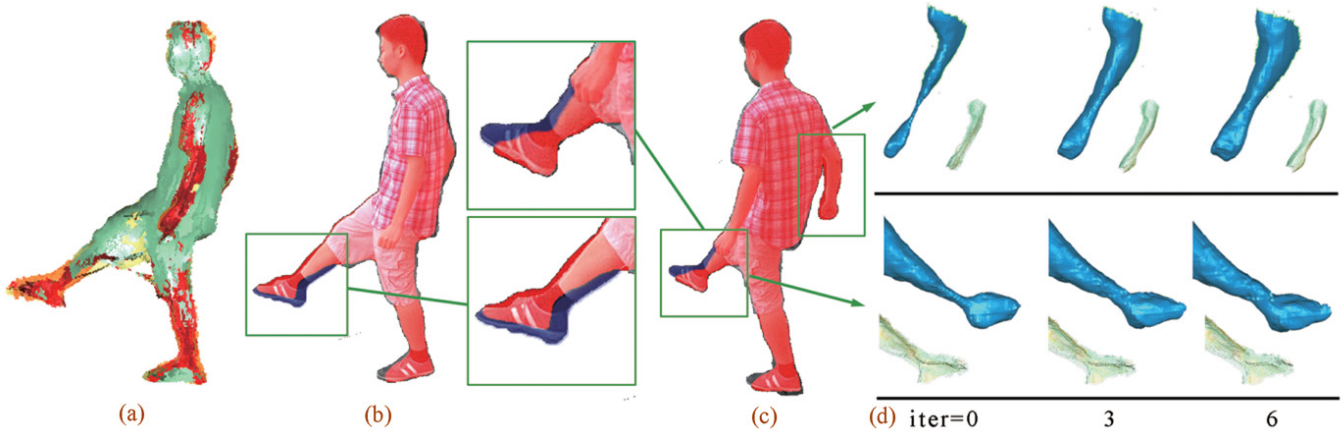


Fig. 4. Point clouds reinforce iteration. (a) Conflict points (red) and normal points (green). (b) and (c) Mesh model into two keyframes to show the leg's movement. The model is projected inside the mask in red regions, and out of the mask in dark red regions. The blue region means there are no projected faces. (d) Two detailed images of the point clouds are extracted to demonstrate the result of the point cloud reinforce iteration. Green part: point clouds with normal vectors. Blue part: result mesh from Poisson reconstruction. For the reason that body shakes and bad matches accumulate, the slender part becomes deformed and even broken. The point's movement will converge when it meets the edge from its vertical plane and form the accordant surface.

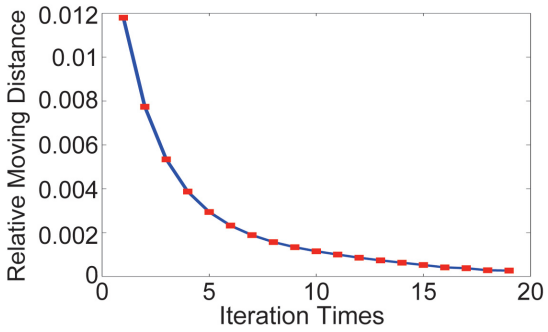


Fig. 5. Convergence of the point cloud reinforce iterations. The vertical coordinate denotes average moving distance of the points. Note that the unit of distance does not equal real-world unit.

where $N(i)$ restricts the hunting zone, the searching radius refers to the size of the model, and '$\cdot$' means the inner product of vectors. α is a tradeoff factor, and in all experiments $\alpha = 0.1$. The point's normal remains invariant during the iteration. The first term is a smooth term, which ensures that the points from one keyframe are continuous, and the other term alters the points' position. $\lambda(i,h)$ forces the points to expand, and $\mu(i,h)$ eliminates the effect of points facing inconformity direction. Because each point is corresponding with a pixel inside the segment masks, the essence is to push the points out of the other frame silhouettes along the normal direction. The silhouette information in each point is changing gently during the iteration to adapt to the other points. Fig. 5 demonstrates the convergence of the iteration.

B. Silhouette Adaption

Since the points reinforcement mainly guarantees the connectivity of the human model, the mesh model is generated from the point clouds using Poisson surface reconstruction [40]. The points reinforcement protects the surface in a low limit by adjusting the points, and more shapes are restored using the silhouette. Following the silhouette adaption method in [17], we constrain the projection of the vertices to lie on the 2D positions on the image silhouette boundary. The refined surface is reconstructed by solving the least squares system

$$\arg\min_{v} \left\{ \|LV - \delta\|_2^2 + \alpha \|C_{\text{sil}} V - q_{\text{sil}}\|_2^2 \right\} \tag{9}$$

where L is the cotangent Laplacian matrix and δ is the differential coordinates of our current mesh with vertices V [43]. α is a weighting factor to the silhouette constraints. The experiment part demonstrates that the method works well in the shaking regions.

C. Close-Shot Refinement

As mentioned in Section IV-A, generally it takes 20–30 s to accomplish the circle around the object human, and this short film is enough to reconstruct an integrated closed model using the above procedure. Beyond that, we could continue to record more close-shot video to improve the details. The close-shot clip means to capture additional video that is closer to the object.

First, the close-shot clip is decoded into image sequence, and is calibrated with the previous images. As the SFM process is run in sequence, additional images are easy to be calibrated with the existing image sequence. To make the SFM robust, the camera should approach the object slowly and maintain a tangential direction moving during the close-shot capture, for the reason that most SFM methods are unstable to forward motion.

Second, we adopt a depth estimation, which is similar to Section III-C. The difference is that both the processing region and the depth searching range are constrained. As the details of the object are magnified in the close-shot image, the texture on a small scale is weakening. Therefore, it is tough to accomplish a fine dense depth estimation, since the textureless area increase and depth estimation will fail in a majority of region. Here, we merely compute the depth in pixels that owns a matched SIFT feature. The SIFT features of each image are pre-extracted in the SFM process. If an

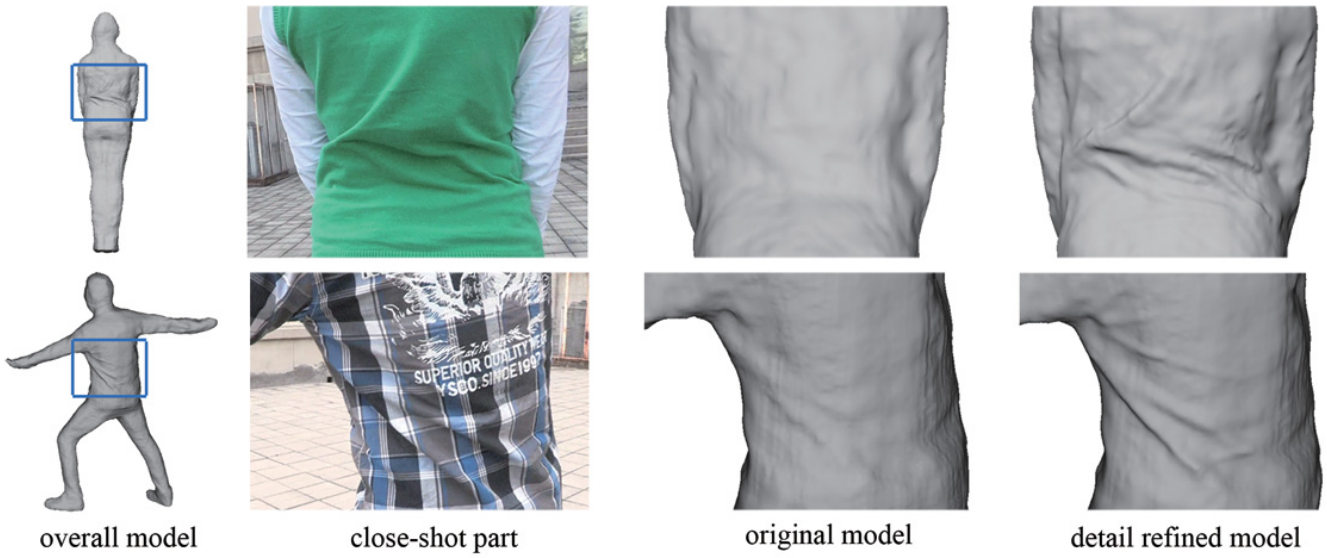


Fig. 6. Close-shot result. In both the refined models, the drapes on the coat are highlighted.

SIFT feature could be matched in more than three images, it is proved to be a reliable feature. The reliable feature positions are recorded after the calibration process, and we use pixels with reliable feature and their 3×3 window neighbor pixels to estimate depth. Furthermore, the depth searching range is constrained in the range that is close to the original mesh model.

In the final step, the vertexes of a pregenerating polygon model are driven by the depth map directly. The pregenerating mesh model is subdivided by simply quartering all of its triangle face. Then, we inverse-project the depth map to the world axis, and generate the points of each pixel. For each pixel, the vertex, which is in the closest distance to the pixel's inv-projecting ray, is judged to be the corresponding vertex. The displacement is transferred from the reliable pixel to the corresponding vertex, and the initial displacement of vertex X, $\Delta p_{\text{ini}}(X)$ is defined as

$$\Delta p_{\text{ini}}(X) = q(X) - p_o(X) \tag{10}$$

where $p_o(X)$ denotes the original position vector and $q(X)$ denotes the corresponding inv-projecting point. Then, the displacement spreads along the topological relation of face. The vertex's displacement $\Delta p(X)$ is formulated as

$$\Delta p(X) = \sum_{i<t} (1 + \cos(i\pi/t))/2 \cdot \Delta p_{\text{ini}}(X_i) \tag{11}$$

where X_i denotes the vertex that is 1-ring neighbor of X, and the final position $p(X)$ is

$$p(X) = p_o(X) + \Delta p(X). \tag{12}$$

In our experiments, $t = 4$. Finally, Laplace smoothing method is employed in the deformed area to ensure that the deformed region could join the undeformed region smoothly. Fig. 6 shows the original model and the close-shot refined result. The final model is rendered with multiview images by the Poisson blending algorithm [42].

V. Experiments

A. Comparison of Results

We compare our approach with the preexisting methods. CMP-MVS [6], 123D Catch,[4] and PhotoScan[5] are image-based 3D modeling softwares, which have been proved to be excellent on the static objects. We capture the video and decode it into image sequence, and then, we use the same image sequence to reconstruct a human body in the aforementioned softwares and our system separately. The video is captured by Cannon XF305 recorder in 1280 × 720, 25 frames/s, and ∼20 s. CMP-MVS, PhotoScan, and our method take 200 frames as input, while 123D Catch takes 70 interval sampling frames as input due to the software input limitation. The image quality is limited due to the unstable illumination and handheld shoot while walking.

Unlike still things, human body is instable and exists tiny shake, which leads to failure in the slender parts like arms, legs, and textureless parts like hair. During the capture of the top sequences in Fig. 7, the object exists obvious shake, and more details could be seen in the supplement videos. As demonstrated in Fig. 7, the red boxes highlight the disconnect or abnormal parts in the CMP-MVS, 123D Catch, and PhotoScan models, while our approach produces more preferable results, protects the slender parts, and retains the details.

B. Quantitative Evaluation

We make the quantitative evaluation of the proposed method and other methods in Fig. 8. The experimental subject is a 1.85 m high static manikin dressed in common clothes [see Fig. 8(a)]. We use a Kinect to scan the full body and reconstruct the mesh model using algorithm [44], which is the modified approach based on Kinect Fusion [28].

[4]http://www.123dapp.com/catch
[5]http://www.agisoft.com

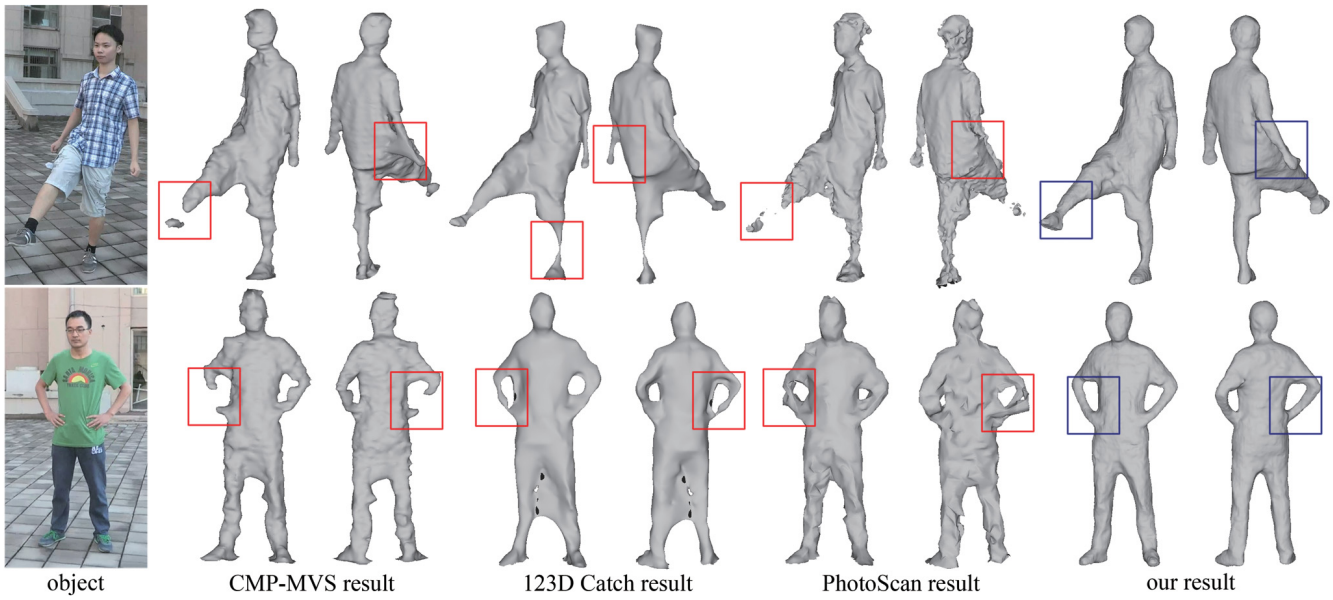


Fig. 7. Comparing with CMP-MVS and PhotoScan, our method protects the shaking and slender parts as shown in rectangle frames.

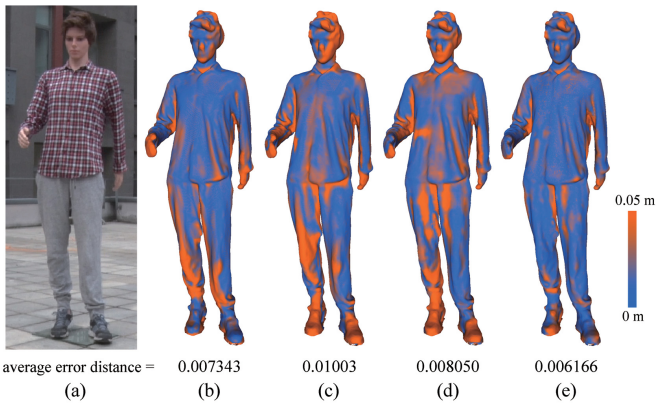


Fig. 8. Quantitative evaluation of different methods. We scanned a manikin (a) with a Kinect, and then evaluated the accuracy of four different methods (b) PhotoScan, (c) 123D Catch, (d) our methods without surface reinforce and refinement, and (e) full pipeline of the proposed method. The average error distances are listed below the model.

Then, the results of PhotoScan, 123D Catch, and our method are compared with the scanned model. When scanning the manikin in the outdoor scenes, we find that the Kinect could hardly work in sunny day due to the interference of the ambient light, and more than 60% pixels of the range image return invalid values. Therefore, we scanned the manikin indoors in advance.

The video is captured by a Sony HXR-NX3 recorder in 1280 × 720, 25 frames/s, and then 255 frames are extracted as the input. Note that 123D Catch takes 70 interval sampling frames as input due to the software input limitation. To make the virtual scale equal to real world scale, we select more than 20 point pairs from the anchor points, and measure the distance of these pairs. The scale factor is the average value of all these pairs' scales, which are defined as the real distance divided by the virtue distance. After zooming the result model to the real world scale, iterative closest point algorithm [45] is used to register the result model to the scanned model. We compute the closest distance between one vertex on the scanned model and the mesh model to be evaluated, and assign the distance value to this vertex, as shown in Fig. 8(b)–(e).

TABLE I
COMPUTATIONAL COST OF THE EXPERIMENT ON DATA SET (b)

Algorithm Step	cost Time/min
Calibration(3.1)	13
Segmentation(3.2)	6.18
Depth Estimation(3.3&3.4)	14
Filter & Normal Estimation	2.58
Point Clouds Reinforce(4.1)	3.25
Silhouette Adapt(4.2)	0.35
Rendering	0.67
Overall	40

The static object reconstruction is steadier than real human body, lacking disconnect problem. Nonetheless, the image demonstrates that the results of PhotoScan [Fig. 8(b)], 123D Catch [Fig. 8(c)] and our pipeline without part III [Fig. 8(d)] degrade in legs and arms; meanwhile, our method [Fig. 8(e)] protects slender parts of the body and generates a more accurate model.

C. Experiments in Various Condition

We verify the proposed method using various challenging real-captured sequences as well. As shown in Fig. 9, other reconstructed 3D model and capture scenes are presented. The capturing equipment includes professional video recorder, commodity DV, and entry-level digital single lens reflex camera. During the procedure of the video shoot, the object distance is ∼3 m. The focal length, aperture size, and white balance are fixed. Fig. 10 shows the 3D printing results using the gypsum powder color 3D printing technique.

Different conditions and challenges are presented in the scenes, respectively: the illuminance in Fig. 9(b) and (c) is

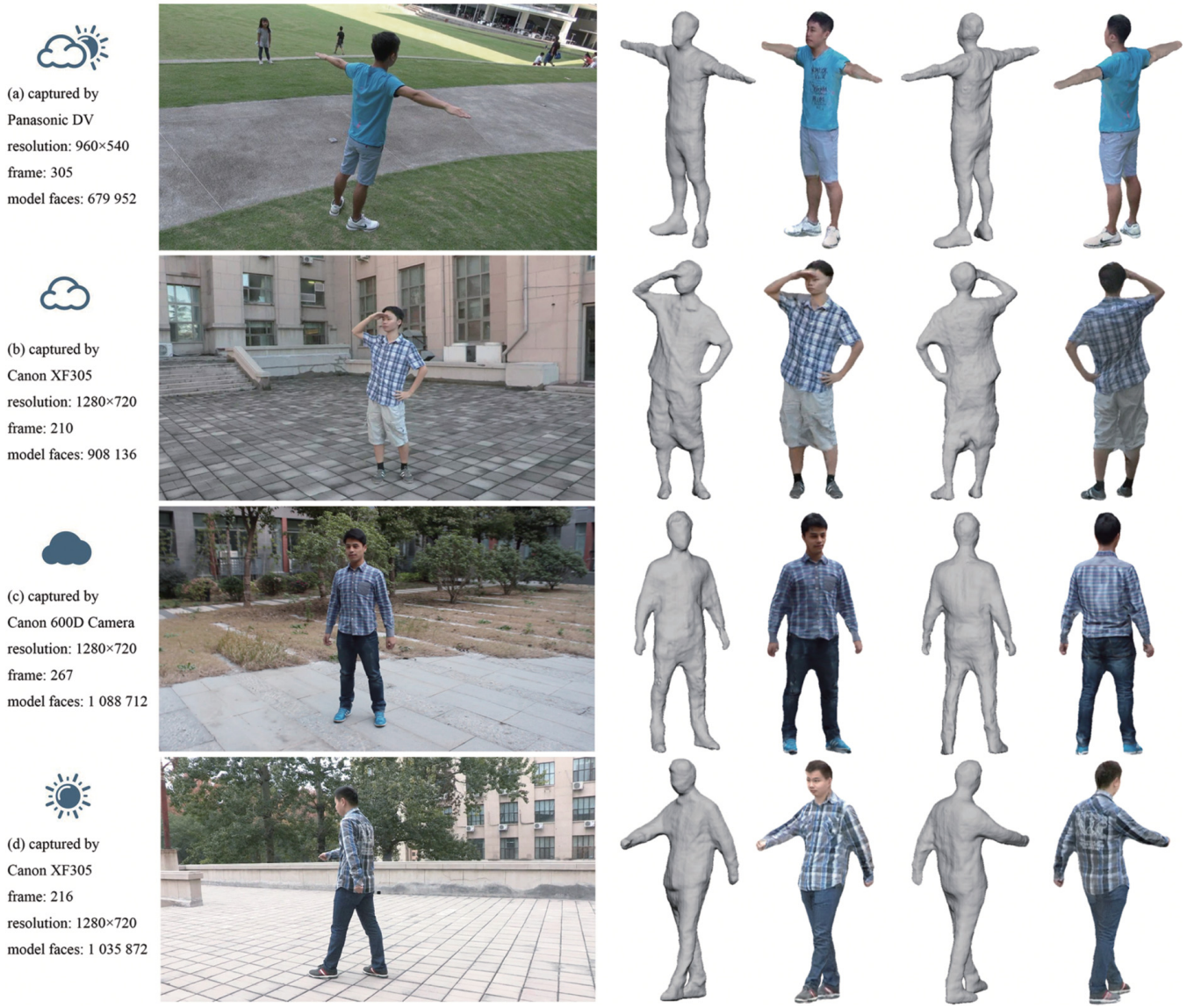


Fig. 9. (a)–(d) Result models captured by our system. For each group, from left to right, we show the video parameters, capture scene, mesh model, and rendering model.

Fig. 10. Our result model is watertight, and can be 3D printed directly.

relatively uniform, since the weather is cloudy, while the scenes in Fig. 9(a) and (d) are exposed under strong sunlight; our method accomplishes the reconstruction in both sunlight and cloudy weather. In Fig. 9(a), there are few people walking in the distance. During the shooting process, the object exists different levels of shake, and more detailed exhibition could be seen in our additional video.

In general, it takes 20 s to orbit the objective figure, and the keyframes are selected out every 15 image, and each cluster has 29 reference frames. We measured the performance of our algorithms on a workstation (CPU i7-2600 3.4 Ghz, 8 Cores, 32-GB RAM). Table I shows the runtime of data set (b). The close-shot refinement is not included in the automatic pipeline, since the close-shot length is uncertain.

VI. Conclusion

We have presented a practical system of scanning a human body using only a conventional video camera. The algorithm performs robustly in outdoor scenes; point clouds reinforcement and silhouette adaption repair the broken regions in legs and arms. Our result models are watertight, which can be

directly used in 3D printing. The overall system is approximately automatic, as only little interaction with user is needed in a segmentation phase.

Our approach still has some limitations. The method cannot handle the case of large-scale movement, and also depends on the success of both the SFM and multiview segmentation algorithms used. In our experiments, most of the video clips did well in SFM, but a few may generate uncorrected calibration due to the motion blur and severe distortion when using the wide field camera.

REFERENCES

[1] X. Lu and X. Liu, "Reconstruction of 3D model based on laser scanning," in *Innovations in 3D Geo Information Systems*. Berlin, Germany: Springer, 2006, pp. 317–332.

[2] J. Geng, "Structured-light 3D surface imaging: A tutorial," *Adv. Opt. Photon.*, vol. 3, no. 2, pp. 128–160, 2011.

[3] J.-S. Park, "Interactive 3D reconstruction from multiple images: A primitive-based approach," *Pattern Recognit. Lett.*, vol. 26, no. 16, pp. 2558–2571, 2005.

[4] C. Wu, "Towards linear-time incremental structure from motion," in *Proc. IC3D*, 2013, pp. 127–134.

[5] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, Aug. 2008.

[6] M. Jancosek and T. Pajdla, "Multi-view reconstruction preserving weakly-supported surfaces," in *Proc. CVPR*, Jun. 2011, pp. 3121–3128.

[7] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. ICCV*, 2011, pp. 2320–2327.

[8] C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen, "On benchmarking camera calibration and multi-view stereo for high resolution imagery," in *Proc. CVPR*, 2008, pp. 1–8.

[9] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. CVPR*, vol. 1. Jun. 2006, pp. 519–528.

[10] Y. Liu, X. Cao, Q. Dai, and W. Xu, "Continuous depth estimation for multi-view stereo," in *Proc. CVPR*, 2009, pp. 2121–2128.

[11] J. Li, E. Li, Y. Chen, L. Xu, and Y. Zhang, "Bundled depth-map merging for multi-view stereo," in *Proc. CVPR*, 2010, pp. 2769–2776.

[12] M. Pollefeys *et al.*, "Visual modeling with a hand-held camera," *Int. J. Comput. Vis.*, vol. 59, no. 3, pp. 207–232, 2004.

[13] M. Pollefeys *et al.* "Detailed real-time urban 3D reconstruction from video," *Int. J. Comput. Vis.*, vol. 78, nos. 2–3, pp. 143–167, 2008.

[14] J. Stühmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *Pattern Recognition* (Lecture Notes in Computer Science), Berlin, Germany: Springer, vol. 6376. 2010, pp. 11–20.

[15] K. Kolev, P. Tanskanen, P. Speciale, and M. Pollefeys, "Turning mobile phones into 3D scanners," in *Proc. CVPR*, 2014, pp. 3946–3953.

[16] Y. Furukawa and J. Ponce, "Carved visual hulls for image-based modeling," in *Proc. ECCV*, 2009, pp. 564–577.

[17] J. Gall, C. Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel, "Motion capture using joint skeleton tracking and surface estimation," in *Proc. CVPR*, 2009, pp. 1746–1753.

[18] D. Cremers and K. Kolev, "Multiview stereo and silhouette consistency via convex functionals over convex domains," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1161–1174, Jun. 2011.

[19] S. Y. Bao, M. Chandraker, Y. Lin, and S. Savarese, "Dense object reconstruction with semantic priors," in *Proc. CVPR*, 2013, pp. 1264–1271.

[20] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Trans. Graph.*, vol. 27, no. 3, 2008, Art. no. 98.

[21] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," in *Proc. SIGGRAPH*, 2008, Art. no. 97.

[22] C. Wu, K. Varanasi, Y. Liu, H.-P. Seidel, and C. Theobalt, "Shading-based dynamic shape refinement from multi-view video under general illumination," in *Proc. ICCV*, 2011, pp. 1108–1115.

[23] C. Wu, Y. Liu, Q. Dai, and B. Wilburn, "Fusing multiview and photometric stereo for 3D reconstruction under uncalibrated illumination," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 8, pp. 1082–1095, Aug. 2011.

[24] M. Pesce, L. M. Galantucci, and F. Lavecchia, "A 12-camera body scanning system based on close-range photogrammetry for precise applications," *Virtual Phys. Prototyping*, vol. 11, no. 1, pp. 49–56, 2016.

[25] K. E. Peyer, M. Morris, and W. I. Sellers, "Subject-specific body segment parameter estimation using 3D photogrammetry with multiple cameras," *PeerJ*, vol. 3, p. e831, Mar. 2015.

[26] M. Pesce, L. M. Galantucci, G. Percoco, and F. Lavecchia, "A low-cost multi camera 3D scanning system for quality measurement of non-static subjects," *Procedia CIRP*, vol. 28, pp. 88–93, Apr. 2015.

[27] S. Bragança, P. M. Arezes, and M. Carvalho, *Occupational Safety and Hygiene III*. Boca Raton, FL, USA: CRC Press, 2015.

[28] R. A. Newcombe *et al.*, "KinectFusion: Real-time dense surface mapping and tracking," in *Proc. ISMAR*, 2011, pp. 127–136.

[29] J. Tong, J. Zhou, L. Liu, Z. Pan, and H. Yan, "Scanning 3D full human bodies using Kinects," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 4, pp. 643–650, Apr. 2012.

[30] W. Chang and M. Zwicker, "Range scan registration using reduced deformable models," *Comput. Graph. Forum*, vol. 28, no. 2, pp. 447–456, 2009.

[31] W. Chang and M. Zwicker, "Global registration of dynamic range scans for articulated model reconstruction," *ACM Trans. Graph.*, vol. 30, no. 3, 2011, Art. no. 26.

[32] H. Li, E. Vouga, A. Gudym, L. Luo, J. T. Barron, and G. Gusev, "3D self-portraits," *ACM Trans. Graph.*, vol. 32, no. 6, 2013, Art. no. 187.

[33] A. Barmpoutis, "Tensor body: Real-time reconstruction of the human body and avatar synthesis from RGB-D," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1347–1356, Oct. 2013.

[34] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, "Multicore bundle adjustment," in *Proc. CVPR*, 2011, pp. 3057–3064.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[36] P. Moulon, P. Monasse, and R. Marlet, "Adaptive structure from motion with *a contrario* model estimation," in *Proc. ACCV*, 2012, pp. 257–270.

[37] W. Lee, W. Woo, and E. Boyer, "Silhouette segmentation in multiple views," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 7, pp. 1429–1441, Jul. 2011.

[38] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: Image and video synthesis using graph cuts," in *Proc. SIGGRAPH*, 2003, pp. 277–286.

[39] T. Beeler, B. Bickel, P. Beardsley, B. Sumner, and M. Gross, "High-quality single-shot capture of facial geometry," in *Proc. SIGGRAPH*, 2010, Art. no. 40.

[40] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Trans. Graph.*, vol. 32, no. 3, 2013, Art. no. 29.

[41] P. Merrell *et al.*, "Real-time visibility-based fusion of depth maps," in *Proc. ICCV*, 2007, pp. 1–8.

[42] M. Chuang, L. Luo, B. J. Brown, S. Rusinkiewicz, and M. Kazhdan, "Estimating the Laplace–Beltrami operator by restricting 3D functions," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1475–1484, 2009.

[43] M. Botsch and O. Sorkine, "On linear variational surface deformation methods," *IEEE Trans. Vis. Comput. Graphics*, vol. 14, no. 1, pp. 213–230, Jan./Feb. 2008.

[44] S. Choi, Q.-Y. Zhou, and V. Koltun, "Robust reconstruction of indoor scenes," in *Proc. CVPR*, 2015, pp. 5556–5565.

[45] S. Rusinkiewicz and M. Levoy, "Efficient variants of the ICP algorithm," in *Proc. 3DMI*, 2001, pp. 145–152.

Hao Zhu received the B.S. degree from the Department of Electronic Science and Technology, Nanjing University, Nanjing, China, in 2013, where he is currently pursuing the Ph.D. degree in electronic science and technology.

His current research interests include computer vision and computational imaging.

Yebin Liu received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2002, and the Ph.D. degree from the Automation Department, Tsinghua University, Beijing, in 2009.

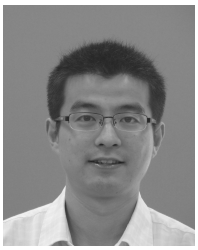
He has been a Research Fellow with the Computer Graphics Group, Max Planck Institute for Informatics, Saarbrücken, Germany, since 2010. He is currently an Associate Professor with Tsinghua University. His current research interests include computer vision and computer graphics.

Jingtao Fan received the B.E. and M.E. degrees in computer science and technology and the Ph.D. degree in optical engineering from the Changchun University of Science and Technology, Changchun, China, in 2003, 2007, and 2013, respectively.

He currently holds a post-doctoral position with Tsinghua University, Beijing, China. His current research interests include 3D video processing and computer vision.

Qionghai Dai (SM'08) received the B.S. degree in mathematics from Shanxi Normal University, Xi'an, China, in 1987, and the M.E. and Ph.D. degrees in computer science and automation from Northeastern University, Shenyang, China, in 1994 and 1996, respectively.

He has been with Tsinghua University, Beijing, China, since 1997, where he is currently a Professor and the Director of the Broadband Networks and Digital Media Laboratory. His current research interests include video communication, computer vision, and graphics.



Xun Cao (M'12) received the B.S. degree from Nanjing University, Nanjing, China, in 2006, and the Ph.D. degree from the Department of Automation, Tsinghua University, Beijing, China, in 2012.

He held visiting positions with Philips Research, Aachen, Germany, in 2008, and Microsoft Research Asia, Beijing, from 2009 to 2010. He was a Visiting Scholar with the University of Texas at Austin, Austin, TX, USA, from 2010 to 2011. He is currently a Professor with the School of Electronic Science and Engineering, Nanjing University. His current research interests include computational photography and image-based modeling and rendering.