3D Body and Background Reconstruction in a Large-scale Indoor Scene using Multiple Depth Cameras

Daisuke Kobayashi
Graduate School of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
kobayashi@limu.ait.kyushu-u.ac.jp

Diego Thomas
Faculty of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
thomas@ait.kyushu-u.ac.jp

Hideaki Uchiyama
Library
Kyushu University
Fukuoka, Japan
uchiyama@limu.ait.kyushu-u.ac.jp

Rin-ichiro Taniguchi
Faculty of Information Science and Electrical Engineering
Kyushu University
Fukuoka, Japan
rin@kyudai.jp

***Abstract*—3D reconstruction of indoor scenes that contain non-rigidly moving human body using depth cameras is a task of extraordinary difficulty. Despite intensive efforts from the researchers in the 3D vision community, existing methods are still limited to reconstruct small scale scenes. This is because of the difficulty to track the camera motion when a target person moves in a totally different direction. Due to the narrow field of view (FoV) of consumer-grade red-green-blue-depth (RGB-D) cameras, a target person (generally put at about $2-3$ meters from the camera) covers most of the FoV of the camera. Therefore, there are not enough features from the static background to track the motion of the camera. In this paper, we propose a system which reconstructs a moving human body and the background of an indoor scene using multiple depth cameras. Our system is composed of three Kinects that are approximately set in the same line and facing the same direction so that their FoV do not overlap (to avoid interference). Owing to this setup, we capture images of a person moving in a large scale indoor scene. The three Kinect cameras are calibrated with a robust method that uses three large non parallel planes. A moving person is detected by using human skeleton information, and is reconstructed separately from the static background. By separating the human body and the background, static 3D reconstruction can be adopted for the static background area while a method specialized for the human body area can be used to reconstruct the 3D model of the moving person. The experimental result shows the performance of proposed system for human body in a large-scale indoor scene.**

***Index Terms*—3D modeling, RGB-D camera, human body, dynamic reconstruction, large scale**

I. Introduction

EpicGame's famous video game Fortnite is a free game where players can pay to personalize their virtual avatar. In 2018, Fortnite reached one millard dollars turnover. There is clearly a huge demand for personalized avatars and several companies like Apple or Huawei (emoji), or start-ups like Meo are now making much efforts to develop systems that can easily generate high definition personal avatars. Outside of the game industry, personal avatars are also used for sports performance analysis and entertainment (free-viewpoint video for example).

As a consequence, how to automatically reconstruct a 3D model of the human body and capture its motion from sensors has drawn much attention in the computer vision research community in the last decades. Thereby, IMUs attached to the body, multi-view imaging and laser scanners have been used for 3D human body reconstruction and motion capture [1]. Recently, leveraging on advances in Red-Green-Blue-Depth (RGB-D) cameras, 3D shape and a human motion can be captured simultaneously when the scene is small and the motion is well controlled.

With the emergence of affordable RGB-D sensors such as the Microsoft Kinect, extensive research has been done on reconstructing 3D models from a video stream of RGB-D images. Reconstructing 3D models of static scenes using volumetric Truncated Signed Distance Function(TSDF) fusion is now a well-understood and matured technology such as the InfiniTAM [2] project, which is readily available and open to the community. Introduced by DynamicFusion [3] in 2015, combining volumetric TSDF with sparse warpfield allows to reconstruct non-rigdly deforming 3D surfaces. However, the large memory and computational cost of these methods preclude from reconstructing large scale scenes. This is particularly a problem for human body reconstruction because when the humans move they move in a large space (ex. a playground). Recently, a technique was proposed that significantly reduced the memory consumption [4] by separating the

978-1-7281-1571-9/19/ $31.00

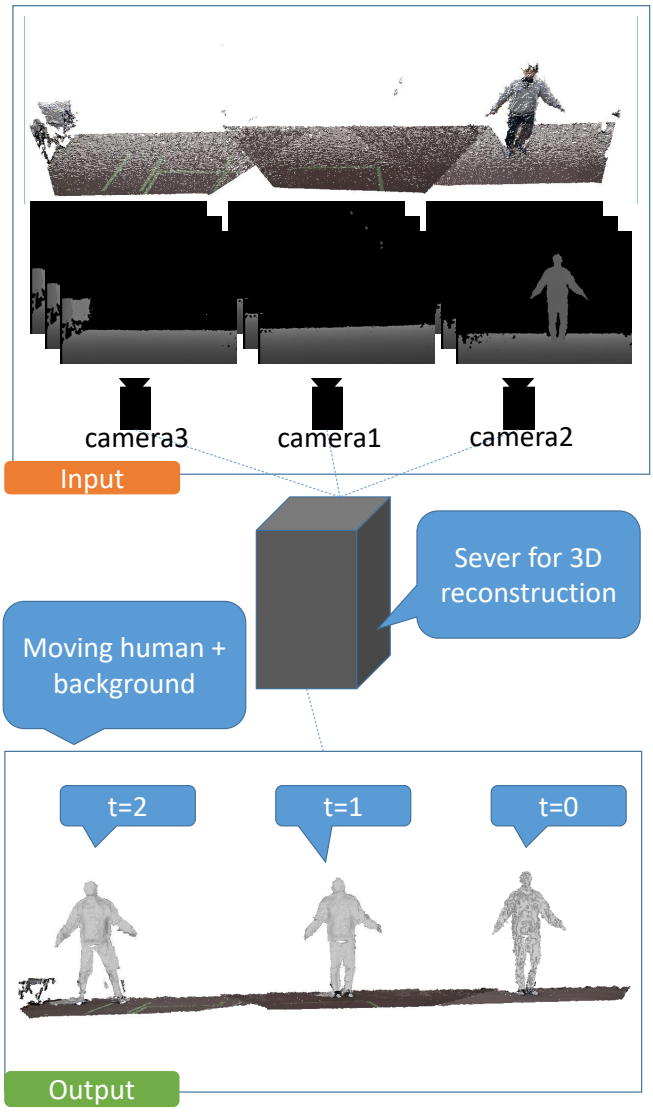


Fig. 1: Overview of our proposed system. We capture the indoor space by three cameras(Top), and reconstruct the 3D model in one computer(middle). We visualize the captured scene(bottom).

problem of reconstructing the full body into multiple independent reconstructions of nearly-rigid body parts. However, no solution has been reported yet that actually reconstructs the human body at large scale.

When the target person moves in a (large) room, the camera must follow the target to continue capturing images and continue the 3D reconstruction process. At that time, the camera motion must be precisely estimated to keep consistent pose of the person with respect to the background. Because of the limited field of view of consumer-grade RGB-D cameras such as the Kinect, the human body covers most part of the image. Therefore, there are only few points on the background that can be used to rigidly align the background part of the 3D shape captured by successive RGB-D images, which makes estimating the camera trajectory impossible. As a consequence, it becomes extremely difficult to reconstruct the 3D human body at large scale.

In this work, we reason that the target person moves inside a predefined large room (like playground for example). Therefore, we can use multiple fixed RGB-D cameras, which added field of view covers the whole area, instead of using a single moving camera. Because all cameras are fixed, the problem of estimating camera trajectory disappears, and we can reconstruct the 3D model of a moving person with the background simultaneously, and at large scale. Now, the main problems that remain are: (a) how to calibrate all cameras in both time and space and (b) how to use the multiple sensors into a unified framework.

In this paper, we propose a system for simultaneous reconstruction of the moving 3D human body and background at large scale by using three non-overlapping Kinect cameras. Our proposed system is composed of three Kinects that are approximately set in the same line and that face in the same direction. The cameras are positioned so that their field of view do not overlap, thus avoiding interference between the depth cameras. The three cameras extrinsic parameters are estimated as two pairs of neighboring cameras. We use three large planar boards that can be seen by two neighboring cameras, and identify the relative transformation matrix that aligns the three normal vectors of the three planes. All cameras are connected to the same computer and synchronized with the internal clock of the computer. To reconstruct the 3D model of the human body, we extend the SegmentedFusion method [4] that builds the model by separating the body into several nearly rigid body parts. The 3D model is initialized with the first set of depth images, then deformation is tracked using the skeleton information provided by the Kinect sensor.

Our contribution in this paper is three-fold: (1) we propose the first system that is able to simultaneously reconstruct the 3D model of a moving person and the background of the scene at large scale; (2) we propose an easy-to-use method to calibrate our the extrinsic parameters of our proposed system; (3) we extend SegmentedFusion to dynamically switch between the different cameras. Fig. 1 shows an overview of our proposed system.

II. Related Work

We separate the related work on 3D reconstruction using RGB-D cameras into two categories: (A) methods for rigid background reconstruction and (B) ones for dynamic human body reconstruction. In this section, we only discuss mostly-related works.

A. Background reconstruction

In the last ten years, consumer-grade RGB-D sensors have become a commodity tool for reconstructing static 3D scenes (such as backgrounds). In 2011, Newcombe et.al. [5] introduced KinectFusion, a real-time system that fuses input

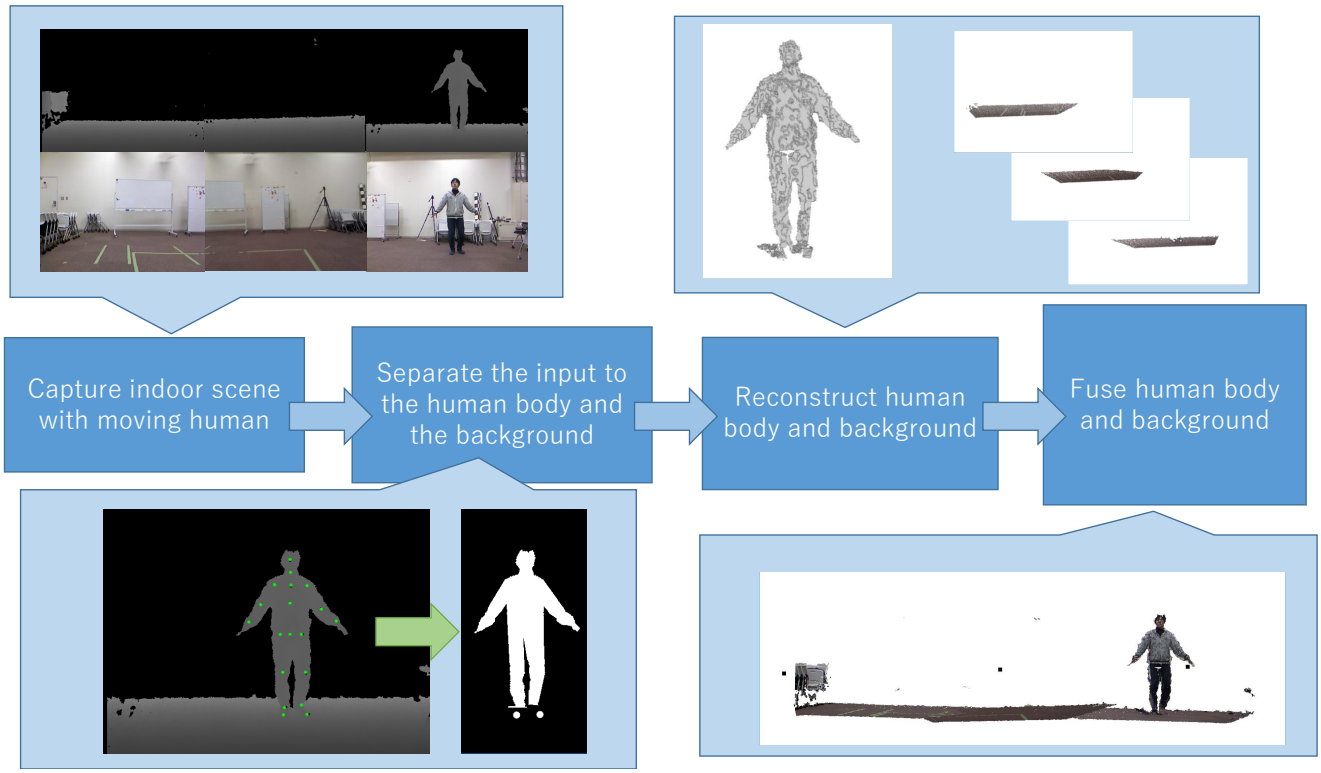


Fig. 2: Pipeline of proposed system. Our system captures the indoor scene by three cameras, then segments to the human body and the background. We separately reconstruct the human body and background, and fuse the reconstructions to visualize the 3D model.

depth images into a volumetric TSDF grid. This was the first successful system that took full advantage of the Kinect sensor. However, its huge memory footprint limits its application to small-scale scenes. Following KinectFusion, a plethora of techniques have been proposed to improve both accuracy scalability of the system. In [6], Nguyen proposed to take into account the model of the sensor noise when fusing depth value to improve accuracy. Roth et. al [7], and Whelan et. al. [8] proposed to move the volume that contains the TSDF field following the camera motion to reconstruct larger-scale scenes. Hornung et.al., [9] and Neissner et.al., [10] proposed to use more compact 3D volumes using Octrees or HashMaps. In [11], the authors proposed to segment the scene into multiple volumes around points of interest to reduce the size of the 3D space to be reconstructed. In [12], Thomas et.al., proposed to encode the geometry over 2D planes detected in the scene to reduce memory consumption as much as possible. In [13] [14], the authors proposed to use planar constraints to improve camera tracking accuracy. Among them, InfiniTAM [2] is a well-developed open-source solution that extends KinceFusion for large scale scene and low computation cost by efficient memory use and voxel hash. However, only static scene can be reconstructed.

B. Human body reconstruction

In 2015, Newcombe et.al. presented DynamicFusion [3], the first system that is capable of reconstructing non-rigid 3D shapes in real-time with a RGB-D camera. The authors proposed to combine the well known TSDF field with a sparse non-rigid warp field. In the DynamicFusion, the 3D surface is sampled to create nodes that contain different 3D transformations. These transformations are blended using dual-quaternions to warp the current estimate of the 3D surface so that it aligns to the input depth image. The same warp field is used to distort the TSDF field and fuse the depth values in real-time. Though impressive results were reported, the proposed method requires slow and control motion, cannot handle topology changes, and is limited to small scale scenes.

After DynamicFusion appeared, several works have been proposed to extend the method to reconstruct the moving human body. Innmann et.al., [15] proposed to use additional color features to improve non-rigid tracking performances. Model-base method and appearance based method have been proposed in [16] [17] [18] [19]. Also, SCAPE is one of the most famous human body models that can be controlled by many parameters. Many works use this model for template-based reconstruction. DoubleFusion is the state-of-the-art method that uses model based human body reconstruction [20]. BodyFusion [21] proposes a skeleton-embedded surface fusion.

Recently, Yao et. al. [4] proposed a method that is cheap in memory consumption to reconstruct the human body. The authors proposed to separate the human body into several body parts and build a small volumetric TSDF volume around each body part (i.e., bounding boxes). By looking at the motion of the skeleton, all the volumes are deformed when the person moves so that the TSDF field of all body parts follows the non-rigid motion without creating any holes or overlap. In [22], the authors also propose to segment the scene into multiple parts and apply different motion to each part, but the volumes were not deformed, and the method failed in reconstructing fast motions.Our proposed method extends SegmentedFusion [4] to simultaneously use three Kinect sensors and reconstruct both the 3D model of the moving person and the background.

III. Overview

We propose to reconstruct the 3D model of a person moving in a large room together with the background using three fixed RGB-D cameras (we use three Microsoft Kinect V1 cameras). Although a single camera has a limited field of view, the combined field of view of the three depth sensors allows for large scale reconstruction. Because the three cameras are fixed, we do not need to estimate the camera motion, which facilitates the reconstruction process and allows us to obtain stable results.

The camera setup of our proposed system is shown in Fig. 1. We set the cameras so that their field of view do not overlap (to avoid the interference of infrared signals) while minimizing the gap between the cameras to limit data loss and enable continuous tracking of the human body. The three Kinect cameras are connected to the same computer and synchronized using the computer internal clock. Finally, we calibrate the extrinsic parameters of the three cameras (i.e., their relative pose) two-by-two using three large planar boards (See section IV-A for details).

We separate the background from the human body and reconstruct their 3D models independently. On one hand, the static background is reconstructed using 3D point cloud. On the other hand, the 3D model of the human body is reconstructed by extending SegmentedFusion [4] to using three cameras (see section IV-B for details). The skeleton provided simultaneously by the three Kinects are merged into a stream of single skeleton so that the tracking and reconstruction of the human body can be done continuously even when the target is moving from the field of view of one camera to another. The reconstructed background and the 3D model of the human body are superimposed for visualization. In Fig. 2, we show the pipeline of our proposed system.

IV. Details of our proposed system

A. Non-overlapping multiple cameras calibration

In order to use the three Kinect cameras into a unified framework, all cameras must first be calibrated. In other words, we need to estimate the relative positions between all cameras. In our system, it is difficult to set a calibration target that is seen by all the cameras. We propose to calibrate the position of the three depth cameras two-by-two, using two pairs of neighboring cameras. To this end, we use three large planar boards that are observed by two neighboring cameras and optimize the relative poses so that all planes align. Fig. 3 illustrates the camera calibration process.

We propose an easy-to-use method to calibrate two neighboring non-overlapping depth cameras. We set three non parallel boards that are viewed from the two cameras (the ground is actually considered as one of the three boards). For each camera, we compute the 3D cloud of points and estimate the three plane equations using RANSAC [23]. Let D_1 and D_2 be the two input depth images of size 640×480 pixels and P_1 and P_2 the corresponding 3D clouds of points. P_k (for $k = 1$ and $k = 2$) is obtained by the inverse perspective projection as follows:

$$P_k(i,j) = \mathbf{K}^{-1} \begin{bmatrix} i \\ j \\ 1 \\ \frac{1}{D_k(i,j)} \end{bmatrix}, \tag{1}$$

where (i, j) are the pixel coordinates (between $([0:640] \times [0:480])$), $\mathbf{K}$ is the 4×4 intrinsic matrix

$$\mathbf{K} = \begin{pmatrix} f_x & 0 & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2}$$

and $\begin{bmatrix} i & j & 1 & \frac{1}{D_k(i,j)} \end{bmatrix}^\top$ are the homogeneous coordinates of the 2D pixel with the depth value $D_k(i, j)$.

The extrinsic calibration problem is the problem of finding the 4×4 3D transformation matrix $\mathbf{T}_{2,1}$ that aligns P_2 to P_1. From P_1 and P_2 and by using RANSAC, we obtain two sets of three planes $\Pi_1 = \left\{\pi_1^1, \quad \pi_2^1, \quad \pi_3^1\right\}$ in P_1 and $\Pi_2 = \left\{\pi_1^2, \quad \pi_2^2, \quad \pi_3^2\right\}$ in P_2 (note that π_i^1 correspond to π_i^2, we obtain these correspondences manually). Each plane is identified by its normal vector $\mathbf{n}_i^j \in \mathbb{R}^3$ and its distance to the origin d_i^j.

We reason that when 2 cameras capture the same set of 3 non parallel planes from different viewpoints, the orthonormal coordinate system formed by the 3 planes in both camera local coordinate system, are related by the same transformation that align the 2 cameras. Therefore, our proposed calibration algorithm estimates the 3D transformation matrix $\mathbf{T}_{2,1}$ that minimizes the following error function.

$$\mathbf{T}_{2,1} = \min_{\mathbf{T}} (\Sigma_{i=1}^{3} (|\mathbf{T}\hat{\mathbf{n}}_i^2 - \hat{\mathbf{n}}_i^1|^2) + |\mathbf{T}[\Omega_2^\top, 1]^\top - [\Omega_1^\top, 1]^\top|^2) \tag{3}$$

where $\hat{\mathbf{n}}_i^j$ is defined as $[\mathbf{n}_i^{j\top}, 0]^\top$, $|.|^2$ means the norm of the vector, and Ω_k is the intersection point of three planes in Π_k.

In our proposed algorithm, we estimate the rotational part of $\mathbf{T}_{2,1}$ and the translational part of $\mathbf{T}_{2,1}$ separately. The rotation part and the translation part are defined as $\mathbf{R}_{\mathbf{T}_{2,1}}$ and $\mathbf{t}_{\mathbf{T}_{2,1}}$.

Firstly, we estimate $\mathbf{R}_{\mathbf{T}_{2,1}}$ by using the orthonormal coordinate system $W_{\Pi^k} = \left\{\mathbf{v}_1^k, \quad \mathbf{v}_2^k, \quad \mathbf{v}_3^k\right\}$ defined by $\Pi^k (k = 1, 2)$.

Firstly, we select 2 planes from Π^k. Then, we generate an orthonormal system using the cross product as follows:

$$\mathbf{v}_1^k = \frac{\mathbf{n}_1^k \times \mathbf{n}_2^k}{|\mathbf{n}_1^k \times \mathbf{n}_2^k|}, \quad (4)$$

$$\mathbf{v}_2^k = \mathbf{n}_2^k, \quad (5)$$

$$\mathbf{v}_3^k = \mathbf{v}_1^k \times \mathbf{v}_2^k. \quad (6)$$

Because W_{Π^1} and W_{Π^2} represent the same coordinate system that from 2 different viewpoints of camera1 and camera2, the transformation that aligns W_{Π^2} to W_{Π^1} also aligns camera2 to camera1. For this reason, $\mathbf{R}_{\mathbf{T}_{2,1}}$ is computed as:

$$\mathbf{R}_{\mathbf{T}_{2,1}} = \begin{pmatrix} \mathbf{v}_1^1 & \mathbf{v}_2^1 & \mathbf{v}_3^1 \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^2 & \mathbf{v}_2^2 & \mathbf{v}_3^2 \end{pmatrix}^{-1}. \quad (7)$$

Secondly, we estimate $\mathbf{t}_{\mathbf{T}_{2,1}}$ by using the intersection point of three planes in W_{Π_k} defined as Ω_k.

$$\mathbf{n}_1^{k\top} \Omega_k = d_1^k, \quad (8)$$

$$\mathbf{n}_2^{k\top} \Omega_k = d_2^k, \quad (9)$$

$$\mathbf{n}_3^{k\top} \Omega_k = d_3^k. \quad (10)$$

These equations can be represented in a matrix form, and the matrix can be inversed because the three planes are not parallel. Therefore, Ω_k is represented as follow:

$$\Omega_k = \begin{pmatrix} \mathbf{n}_1^{k\top} \\ \mathbf{n}_2^{k\top} \\ \mathbf{n}_3^{k\top} \end{pmatrix}^{-1} \begin{pmatrix} d_1^k \\ d_2^k \\ d_3^k \end{pmatrix}. \quad (11)$$

Similar to $\mathbf{R}_{\mathbf{T}_{2,1}}$, the translation between Ω_1 and Ω_2 is also the transformation between camera1 and camera2. For this reason, $\mathbf{t}_{\mathbf{T}_{2,1}}$ is computed as:

$$\mathbf{t}_{\mathbf{T}_{2,1}} = \Omega_1 - \mathbf{R}_{\mathbf{T}_{2,1}} \Omega_2. \quad (12)$$

As a result, we obtain the rotational part and the translation part of the 3D transformation matrix $\mathbf{T}_{2,1}$. Overall, the 4×4 3D transformation matrix $\mathbf{T}_{2,1}$ is represented as follows:

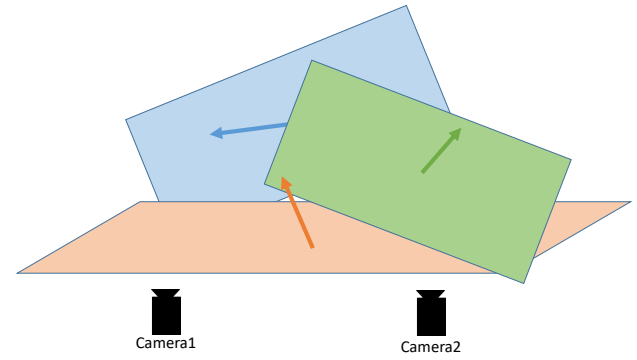
$$\mathbf{T}_{2,1} = \begin{pmatrix} \mathbf{R}_{\mathbf{T}_{2,1}} & \mathbf{t}_{\mathbf{T}_{2,1}} \\ \mathbf{0}^\top & 1 \end{pmatrix}. \quad (13)$$

In the same way, we compute $\mathbf{T}_{3,1}$. As a result, we know all relative positions of the cameras, and our system is calibrated. In Fig. 5, we show the result of the calibration.

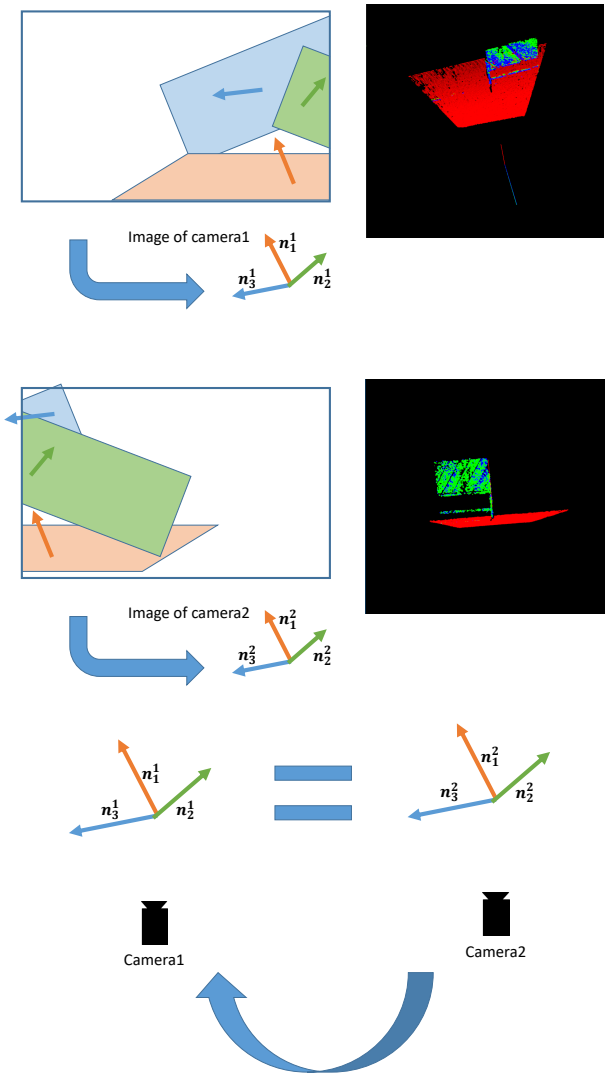
B. Human body reconstruction

We detect and segment out the human body using the skeleton provided by the Kinect sensors. Then, we reconstruct the non-rigidly moving 3D model of the target person by extending SegmentedFusion [4] to using three cameras. The main idea of SegmentedFusion is to reconstruct each body part independently with TSDF fusion, and to non-rigidly stitch all body parts together by looking at the motion of the skeleton. Fig. 4 illustrates the pipeline of the human body 3D reconstruction algorithm.

The main difficulty of extending SegmentedFusion to using three cameras is to carefully handle situations when the target



(a) Capture 3 planes by Kinects



(b) Compute the transformation $\mathbf{T}_{2,1}$

Fig. 3: Illustration of the camera calibration. We capture three planes by Kinects (a), then, we generate the coordinate system and compute the transformation $\mathbf{T}_{2,1}$ (b).

person is in between the field of view of two neighboring cameras. In this situation, the skeleton provided by different Kinect sensors must be merged into a single consistent skeleton to enable continuous and smooth tracking of the person's motion.

First, we briefly recall the human body reconstruction method proposed in [4] and illustrated in Fig. 4. The human body is segmented into multiple (nearly rigid) body parts, and the 3D model is initialized with the first depth image (top of Fig. 4). From the first depth image, the center and bounding box of the each body part are identified using PCA. Then, each bounding box is deformed to tightly stitch all body parts together so that neither holes nor overlap appear. More precisely, a dual quaternion (that represents a 3D transformation) is attached to each joint of the skeleton, and the dual quaternions, that correspond to the two joints that form each bone of each body parts, are linearly blended to deform the 3D shape (of each body part). For each joint, the dual quaternion is estimated from the relative transformation between the child bone and the parent bone (middle of Fig. 4).

At every new input depth image, the dual quaternions of each joint are estimated, which defines the non-rigid warp of the reconstructed 3D surface. Then, new depth value is accumulated into the TSDF field of each body part using the current warp field. Namely, for each body part with joints $\{ch, pa\}$ and corresponding dual quaternions $\{q_{ch}, q_{pa}\}$, and for each voxel v with 3D coordinates $\begin{bmatrix} x & y & z \end{bmatrix}^\top$, we obtain the corresponding pixel index (i, j) in the depth image as follows:

$$(i, j) = K\mathrm{DQB}(v, q_{ch}, q_{pa}), \tag{14}$$

where DQB is the dual quaternion blending function as described in [24] and [4], and K is the intrinsic matrix of the camera. The TSDF value of the voxel v is then updated using the depth value $D(i, j)$ and the standard running average procedure as described in [5].

Because of the noise in the depth image and in the skeleton joints provided by the Kinect sensors, outliers often appear in the reconstructed 3D model. To remove these outliers we check the position of the joint. When the 3D joint position exists out of the human body bounding box, we assume that the joint is outlier. These outliers are not used in skeleton motion estimation.

When the target person is in between the field of view of two adjacent cameras, our proposed system captures two incomplete skeletons. We propose to unify the skeletons provided by multiple Kinects. More precisely, when multiple skeletons are detected for the same person, our system uses all these skeletons to reconstruct the human body. When the joint exists only one, we use the point for the reconstruction, but when the joint exists in several images we select to one point randomly, and run the fusion of the input image twice to update the TSDF volume.

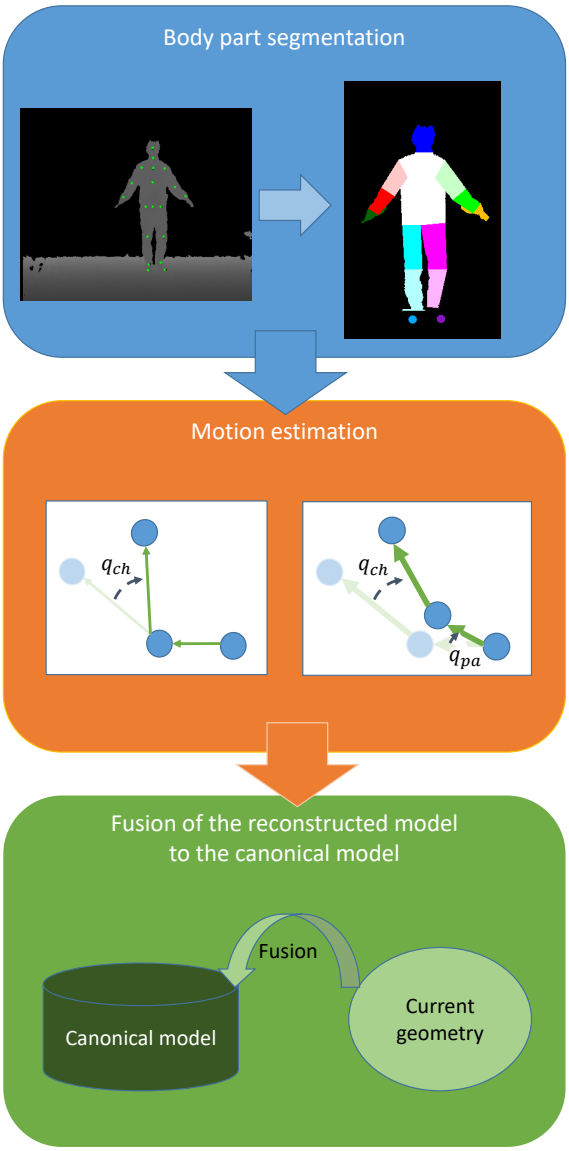


Fig. 4: Pipeline about 3D reconstruction of human body. We segment the human body to the body parts(Top), and estimate the human motion by the 3D joint position(middle). Then, we fuse the current geometry to the canonical model(bottom).

C. Scene reconstruction

As explained in section IV-B, we have separated the background and the human body at every frame. As a consequence, we can use parts of the input images that correspond to the background to reconstruct a static 3D model of the scene.

Up to now, we obtained two components: the background and the human body for the reconstruction of the human body in large scale scenes. From this, we describe the fusion and visualization of these components. On one hand, the human body is visualized by a 3D mesh. In our system, we use the MarchingCubes algorithm [25] to generate the 3D mesh. On the other hand, the static background is visualized by the point

cloud made from the input images. Then, these components are aligned by the result of the calibration. Finally, we visualize the reconstruction of the scene containing the moving person.

V. Experiment

A. Settings

We describe the ability of our proposed system to reconstruct the 3D model of a person moving in a large-scale indoor scene. Our system setup is shown in Fig. 1. We used 3 Kinect V1 cameras the field of view does not overlapped to avoid interference of infrared light while capturing large scale experimental space. In this setup, the distance between Kinects is 3m and the space captured is $10 \times 4 \times 5(w \times h \times d)m^3$. We use the dataset captured in this experimental space.

B. Result

In Fig. 5, we show the results of the calibration process for our camera setup. In this experiment, we check the ability of our calibration method and prove how easy our method can be used. For simplicity, we used 2 images that captured the calibration target (three planar boards): one image for each camera. The upper part of Fig. 5 shows the calibration result obtained for a pair of two neighboring cameras. We can observe that the floor is flat and continuous. The planar boards are also aligned in a common coordinate system. The bottom part of Fig. 5 shows the calibration results after combining results for all three cameras. Same as the results obtained for the pairs of neighboring cameras, all planar boards are well aligned. Although we used a few images for the calibration, the obtained results show that our proposed method could successfully calibrate these cameras.

Fig. 6 shows the reconstruction results obtained for the sequence called "moving human". In this sequence, a person moves about 10m in the capture space of our system. This sequence contains a large-scale background, a moving human body and the switching between cameras. In the first frame, the surface of the human body is rough. As time goes by, the position and pose of the human change and the surface becomes smooth. These results demonstrate the ability of our proposed system for reconstructing moving human body in large scale indoor scene.

C. Discussion

While our system can handle large-scale indoor scenes where a person moves, some limitations exist. First, Our proposed system cannot deal with the twist motion, because the skeleton tracking cannot explain a twist motion and estimated position of the joints is blurred. For solving this problem, we consider that we insert a rigid or non-rigid ICP algorithm to the human reconstruction step. Second, although our system can handle scenes where the person moves between cameras, it is difficult for our system to handle more complex situations such as multiple people the intersect each other. We consider that using the color information for the human identification is one possibility to solve this problem. Third, if the human body is largely occluded, it is difficult to reconstruct the 3D model because of the failure of the skeleton tracking. To solve this problem, we have to use a more robust skeleton tracking method. In the future, we will consider more general scenes such as targeting multiple actions and multiple persons.

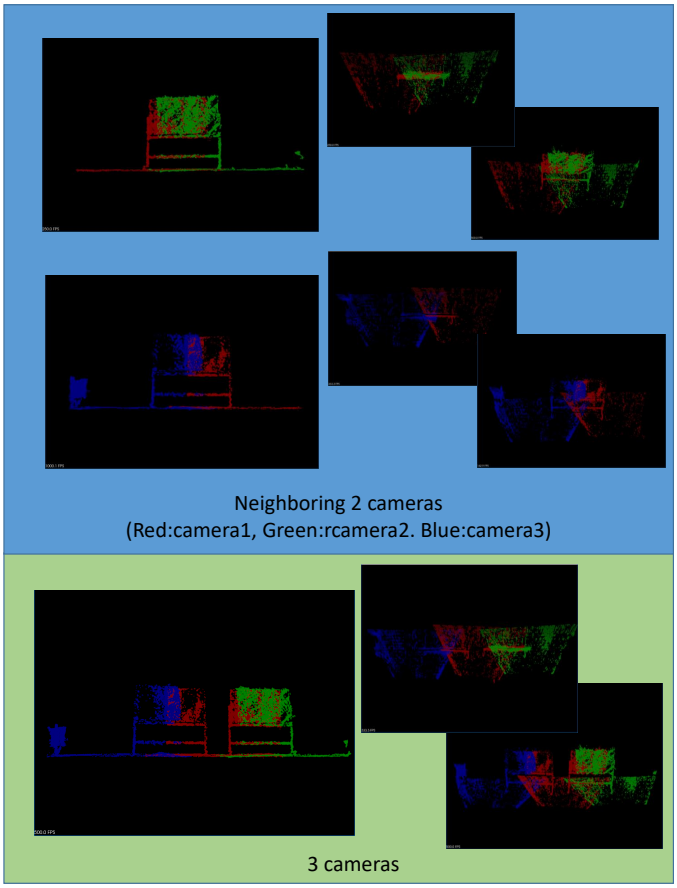


Fig. 5: Result of calibration by our proposed algorithm. The upper part is calibrated by neighboring 2 cameras, the bottom part is merged all of the calibration results.

VI. Conclusion

In this paper, we proposed a system that is able to simultaneously reconstruct the 3D model of a moving person and the background of the scene at large scale. Our system is composed of 3 non-overlapping Kinects which are calibrated using our proposed easy-to-use method. The calibration experiment shows that our algorithm can calibrate the three non-overlapping Kinects. The "moving human" experiment showed that the system is capable of reconstructing the 3D model of a human moving in a large-scale indoor space.

We hope that our work helps future research on 3D modeling of a moving human body, and we believe that it brings the state-of-the-art one step closer to non-rigid 3D reconstruction of large scale scenes.

References

[1] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human

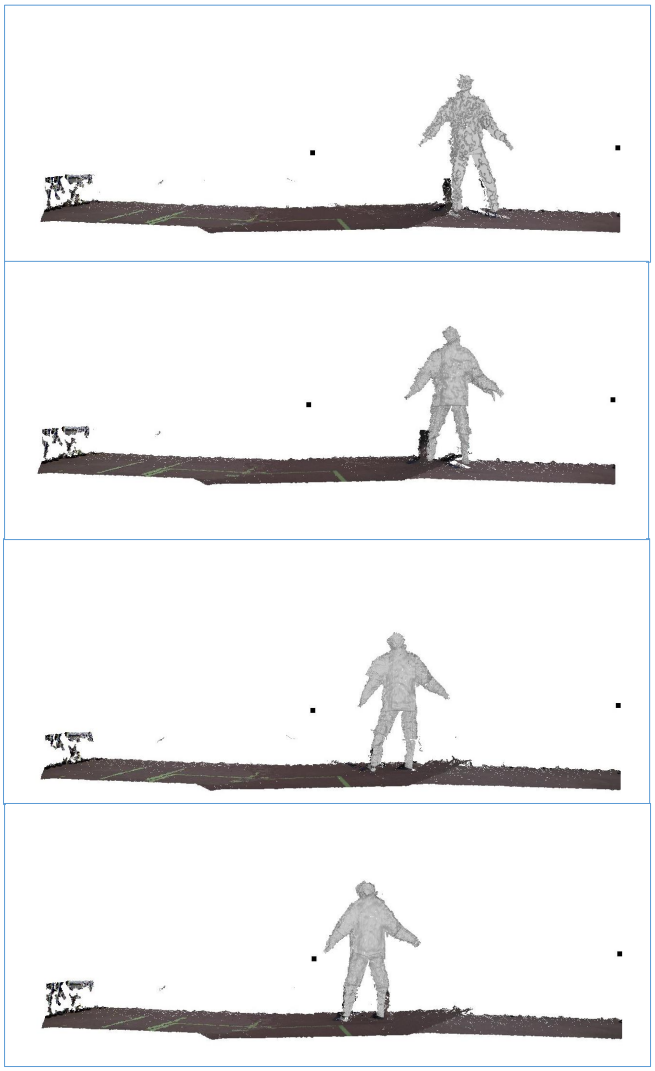


Fig. 6: Reconstruction result of "moving human" scene. The human moves right to left, and the human surface becomes smooth over time.

sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[2] O. Kahler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. H. S Torr, and D. W. Murray. Very High Frame Rate Volumetric Integration of Depth Images on Mobile Device. *IEEE Transactions on Visualization and Computer Graphics*, 22(11), 2015.

[3] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[4] Shih Hsuan Yao, Diego Thomas, Akihiro Sugimoto, Shang-Hong Lai, and Rin-Ichiro Taniguchi Kyushu. Segmentedfusion: 3d human body reconstruction using stitched bounding boxes. In *2018 International Conference on 3D Vision (3DV)*, pages 190–198. IEEE, 2018.

[5] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[6] Chuong V Nguyen, Shahram Izadi, and David Lovell. Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 524–530. IEEE, 2012.

[7] Henry Roth and Marsette Vona. Moving volume kinectfusion. In *BMVC*, volume 20, pages 1–11, 2012.

[8] Thomas Whelan, Michael Kaess, Maurice Fallon, Hordur Johannsson, John Leonard, and John McDonald. Kintinuous: Spatially extended kinectfusion. 2012.

[9] Armin Hornung, Kai M Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. Octomap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous Robots*, 34(3):189–206, 2013.

[10] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):169, 2013.

[11] Qian-Yi Zhou and Vladlen Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics (ToG)*, 32(4):112, 2013.

[12] Diego Thomas and Akihiro Sugimoto. Modeling large-scale indoor scenes with rigid fragments using rgb-d cameras. *Computer Vision and Image Understanding*, 11 2016.

[13] Renato F Salas-Moreno, Ben Glocken, Paul HJ Kelly, and Andrew J Davison. Dense planar slam. In *Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on*, pages 157–164. IEEE, 2014.

[14] Yuichi Taguchi, Yong-Dian Jian, Srikumar Ramalingam, and Chen Feng. Point-plane slam for hand-held 3d sensors. In *ICRA*, pages 5182–5189, 2013.

[15] Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pages 362–379. Springer, 2016.

[16] Federica Bogo, Michael J Black, Matthew Loper, and Javier 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.

[17] Yin Chen, Zhi-Quan Cheng, Chao Lai, Ralph R Martin, and Gang Dang. Realtime reconstruction of an animating human body from a single depth camera. *IEEE transactions on visualization and computer graphics*, 22(8):2000–2011, 2016.

[18] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[19] Chao Zhang, Sergi Pujades, Michael J Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3d scan sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4191–4200, 2017.

[20] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. *arXiv preprint arXiv:1804.06023*, 2018.

[21] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. ACM, October 2017.

[22] Martin Rünz and Lourdes Agapito. Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. *arXiv preprint arXiv:1706.06629*, 2017.

[23] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[24] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Geometric skinning with approximate dual quaternion blending. *ACM Transactions on Graphics (TOG)*, 27(4):105, 2008.

[25] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *ACM siggraph computer graphics*, volume 21, pages 163–169. ACM, 1987.