Immersive Learning Experiences for Surgical Procedures

Young-Woon CHA [a,1], Mingsong DOU [c], Rohan CHABRA [a], Federico MENOZZI [a], Andrei STATE [a,d], Eric WALLEN, MD [b] and Henry FUCHS [a]

[a] *Department of Computer Science, University of North Carolina at Chapel Hill*
[b] *Department of Urology, University of North Carolina at Chapel Hill*
[c] *Microsoft Research*
[d] *InnerOptic Technology, Inc.*

Abstract. This paper introduces a computer-based system that is designed to record a surgical procedure with multiple depth cameras and reconstruct in three dimensions the dynamic geometry of the actions and events that occur during the procedure. The resulting 3D-plus-time data takes the form of dynamic, textured geometry and can be immersively examined at a later time; equipped with a Virtual Reality headset such as Oculus Rift DK2, a user can walk around the reconstruction of the procedure room while controlling playback of the recorded surgical procedure with simple VCR-like controls (play, pause, rewind, fast forward). The reconstruction can be annotated in space and time to provide more information of the scene to users. We expect such a system to be useful in applications such as training of medical students and nurses.

Keywords. 3D Reconstruction, Immersive Experience, Virtual Reality, Surgical Procedures, Immersive Environment

1. Introduction

In this paper, we propose a system for 3D-plus-time recording of activities, such as surgical procedures, through 3D capture and reconstruction methods. We introduce that pro-

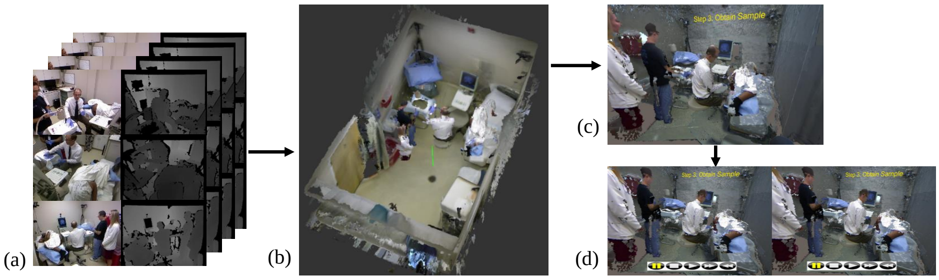


Figure 1. System Pipeline: (a) Recording: Multiple-view RGB-depth image sequences are captured during the procedure. (b) 3D reconstruction: The 3D scene is reconstructed using the sequences. (c) Annotation: The reconstruction is annotated with 3D text for playback. (d) Immersive Experience: A user walks around the reconstruction using a head-mounted display.

[1] Young-Woon Cha – youngcha@cs.unc.edu

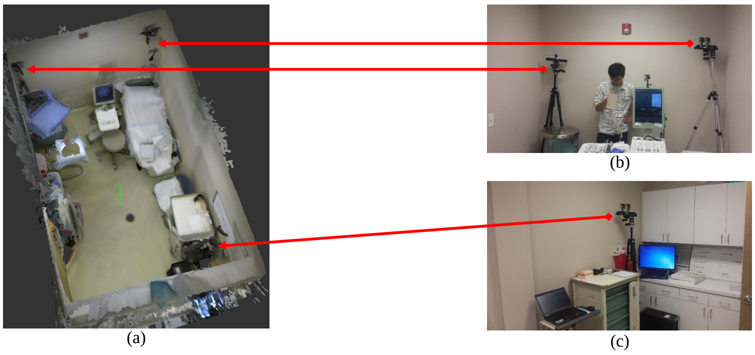


Figure 2. Recording configuration. (a): reconstructed 3D procedure room using depth images from a single moving hand-held camera. (b) and (c): fixed wall-mounted Kinect depth cameras that capture moving objects during the procedure.

vides an application immersive environments for medical training. In our initial prototype shown here, we captured a prostate biopsy procedure at a UNC Urology clinic (Fig. 1). Our system performs dynamic reconstruction for all persons present: patient, physician, nurse assistant, and an observer. The small procedure room was instrumented with three Kinect color+depth cameras in three of its corners. Because of the setup's limited coverage, and frequent occlusion events caused by the participants, this first reconstruction contains spatial gaps and other inaccuracies.

Yet despite its shortcomings, compared with being physically present at the procedure, the virtual presence provided by our prototype system has several advantages: a student experiencing the immersive reconstruction can freely move to any desired viewing location, including locations that might have interfered with the procedure as it was being executed; the reconstruction can be annotated in space and time with information that facilitates insight and accumulation of knowledge–annotations can be added post-reconstruction by the physician who performed the procedure, or by other competent personnel; finally, the student may pause, rewind, or temporally scan through the procedure at variable speed forward or backward in time, or even ”single-step” through it.

The remainder of this paper is organized as follows. After reviewing relevant previous work in Section 2, the proposed framework is detailed in Section 3, which includes descriptions of the dynamic scene reconstruction, annotation, playback, and visualization based on a head-mounted display (HMD). The experimental results are discussed in Section 4. We conclude and summarize possible improvements in Section 5.

2. Background

The modern medical simulator systems [1,2,3] investigate animatable simulators on immersive virtual environments to provide better understanding to users using 3D visualizations than fixed 2D video streams. Limited perspectives are provided to users for immersive visualization. Recent studies [4,5,6] investigate walk-around VR systems using HMD, though these approaches still use predefined meshes or manually reconstruct virtual scenes using 3D graphics tools for real world scenarios.

The immersive environments can be generated directly from recorded images using 3D reconstruction methods for more realistic visualization. In controlled setting, [7] proposes the 3D reconstruction of environments from images with HMD visualization. [8]

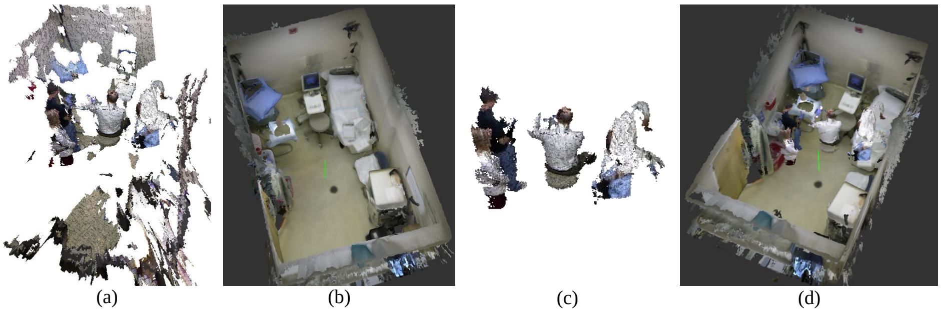


Figure 3. Dynamic scene generation. (a): 3D surface meshes from three Kinects at time t. (b): pre-scanned 3D surface mesh of procedure room. (c): segmented surface mesh from (a). (d): combined mesh consisting of (b) and (c).

shows reconstruction of objects combined with predefined environments using real-time stereo matching. In our new system, the entire immersive environment is fully reconstructed from captured images.

The immersive 3D virtual reality (VR) system we introduce here is similar to previously described telepresence systems [9] and enables users to experience immersive 3D environments through a combination of 3D scanning and immersive display. The 3D scanning methods for dynamic scenes reconstruct a sequence of surfaces by updating changes in the scene over time [10]. Geometric change detection methods [11,12] estimate the changed areas in the scene by modeling static backgrounds. The dynamic 3D scene is updated by re-scanning the changing regions while leaving other regions untouched.

3. Methods

In this section, we detail the proposed approach for generating the immersive learning environments. The system pipeline is illustrated in Fig. 1. First, synchronized multiple-viewpoint RGB-depth image sequences are captured during the procedure, using Microsoft Kinect depth cameras. Second, the entire procedure is reconstructed as a sequence of 3D surface meshes over time, using previously described methods [10]. Third, the sequence of 3D surfaces can be manually annotated by adding timed 3D text in appropriate locations to describe and explain the activities. After this processing, a user wearing a tracked HMD can examine the reconstructed, annotated immersive environment at leisure and repeatedly, as described above. By updating the eye positions provided by the HMD tracker in real-time, the visualization subsystem presents a walkable immersive environment from the user's perspective. The user controls the playback of the reconstruction with a remote hand-held controller such as a wireless mouse. In the following subsections, we elaborate on each step.

3.1. Capture and Dynamic Scene Reconstruction

This subsection describes how surgical procedure scenes are reconstructed as a sequence of surface meshes: $\mathbf{M} = \{\mathbf{M}_1, ..., \mathbf{M}_T\}$. To capture dynamically changing indoor environments over time, we capture the static background (e.g., the room where the procedure takes place) in advance, and acquire dynamically changing objects separately, to handle changes in the surface mesh [10].

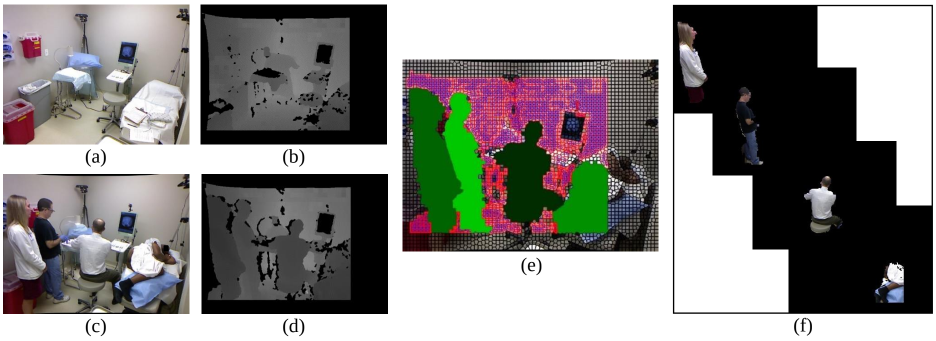


Figure 4. Segmentation of dynamic elements. (a) and (b) are a pair of color and depth images of the empty procedure room. (c) and (d) show the RGB-depth image at time t. From (b) and (d), changed parts (green) are segmented from background (purple) using superpixel based foreground detection in (e). (f) shows the separated segments in (e).

The static background, denoted as $\mathbf{M}_0$, is pre-scanned with a single moving camera (Fig. 2a). We utilized an extended version of KinectFusion [13] for room-sized scene reconstruction that incorporates plane matching to improve reconstructions of features in walls, ceiling, and floor [14].

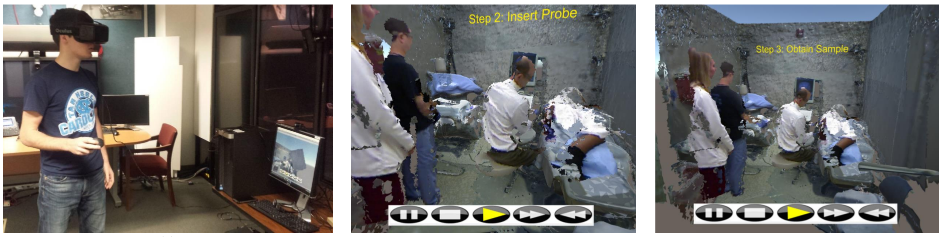
The moving objects (typically, people and instruments) are captured over time by fixed depth cameras mounted in the corners of the room (Fig. 2). The depth cameras are pre-calibrated to a global coordinate system [10]; One of the cameras, C^1, is located at the origin $[\mathbf{I}_{3\times3}|\mathbf{0}_{3\times1}]$ of the global coordinate system, and other cameras, C^i, are at their respective poses $[\mathbf{R}^i_{3\times3}|\mathbf{T}^i_{3\times1}]$ relative to C^1. Let $\mathbf{V}_t = \mathbf{V}^1_t \cup ... \cup \mathbf{V}^N_t$ be a set of colored 3D vertices extracted from RGB-depth images of the N calibrated and synchronized cameras at time t.

The pre-scan $\mathbf{M}_0$ of the static background (Fig. 3b) is also aligned to the global coordinates of the camera cluster. To achieve that, we initially use pose estimation based on SIFT feature matching; then we refine the alignment through ICP registration between $\mathbf{M}_0$ and the initially (at time step 0) acquired live geometry set $\mathbf{V}_0$ [10].

At each subsequent time step, the acquired live geometry set $\mathbf{V}_t$ as shown in Fig. 3a is segmented to detect foreground (i.e. non-background) data $\mathbf{F}_t \equiv \{v | v \in \mathbf{V}_t \text{ and } v \notin \mathbf{M}_0\}$ (Fig. 3c) by comparing $\mathbf{V}_t$ (Fig. 3a) with $\mathbf{M}_0$ (Fig. 3b). The reconstructed surface mesh at frame t is defined as $\mathbf{M}_t \equiv \mathbf{F}_t \cup \mathbf{M}_0$ (Fig. 3d).

The foreground vertices $\mathbf{F}^i_t$ at each camera i are estimated from $\mathbf{V}^i_t$ via superpixel based background subtraction. Fig. 4 shows an example of such foreground segmentation. The static background model $\mathbf{B}^i_0$ at each camera i is estimated from a set of depth images captured just before the procedure (Fig. 4b). The $\mathbf{V}^i_t$ is labeled as $l(v \in \mathbf{V}_t) \in \{0 = background, 1 = foreground\}$ by subtracting $\mathbf{B}^i_0$ from the depth image $\mathbf{D}^i_t$. The color image $\mathbf{I}^i_t$ is segmented as a set of superpixels $\mathbf{S}$ using SLIC [15] by merging local pixels based on the color similarity. (Fig. 4e). The superpixel $S_i \in \mathbf{S}$ includes a set of vertices $v_s \in S_i$, and is labeled as $l(S_i)$ by voting $l(v_s)$. Superpixel-level connected components are extracted based on the similarity of depth values between adjacent superpixels (Fig. 4e-f). The foreground vertices $\mathbf{F}^i_t = \{v | v \in \mathbf{V}^i_t, v \in S_i \text{ and } l(S_i) = 1\}$.

The $\mathbf{F}_t = \mathbf{F}^1_t \cup ... \cup \mathbf{F}^N_t$ represent the colored 3D points that differ from the static background mesh $\mathbf{M}_0$. The $\mathbf{F}_t$ are meshed using the marching cubes algorithm [16] followed by a volumetric fusion pass [17]. This forms the dynamic surface at time t as shown in Fig. 3c. The complete 3D surface mesh becomes $\mathbf{M}_t = \mathbf{F}_t \cup \mathbf{M}_0$ shown in



(a) (b) (c)

Figure 5. Interacting with the immersive reconstruction. (a): The user examines the reconstruction through a head-mounted display. The user controls the playback of the scene with a wireless hand-held controller. (b) and (c): user's views at two different times. The VCR controls are shown at the bottom. Annotations are visible as yellow text on the wall.

Fig. 3d. At visualization time, we render the sequence of dynamic 3D surface meshes $\mathbf{M} = \{\mathbf{M}_1, ..., \mathbf{M}_T\}$ to the user's HMD in real-time.

3.2. Scene Annotation and Playback control

In addition to the sequence of dynamic 3D surface meshes **M**, the user is able to view additional descriptions about the scene and to control the playback of the sequence as mentioned.

To insert the annotations, a subset of the frames $\mathbf{M}_s \equiv \{\mathbf{M}_{t_1}, ..., \mathbf{M}_{t_2}\}$ where $t_1 < t_2$ are manually enhanced with 3D text placed in specific locations in M_s. An example of such annotation is shown in Fig. 5. The text in this example is positioned on the wall in 3D space and provides information about the surgery step occurring during this period.

During playback, the user can quickly move to a specific time period in the recording using the virtual playback controller and a wireless mouse (Figs. 5b and 5c). The controller includes play, pause, stop, fast forward, and rewind buttons.

3.3. Head-mounted Display Visualization

Fig. 6 shows a user walking through the reconstructed procedure room. Using the 3D positions of the user's eyes and the HMD viewing direction supplied by the HMD tracker, the immersive, annotated environment is rendered stereoscopically, distortion-corrected and displayed in the user's HMD.

To enable the user to walk along the floor in the reconstructed procedure room as he or she walks in the real world, the coordinates between the reconstructed scene and of the HMD tracker must be aligned. To accomplish this, we manually transform the reconstructed mesh to align the floor plane ($y = 0$) in the mesh with the floor plane ($y = \alpha$) in the user's room.

The mesh is manually transformed in advance so that the floor in the mesh is located at $y = 0$ in the coordinate. The $x - z$ plane of the HMD tracker is aligned with the ground regardless of camera orientation. The floor planes in the reconstruction and in the real world are aligned by manually adjusting the y coordinate of the mesh.

4. Results

In the recording, four people were present during the procedure shown: a patient, a physician, a nurse assistant, and an observer. To record the scene, four calibrated Microsoft Kinect depth cameras were used; one mobile unit for the pre-scan of the static back-

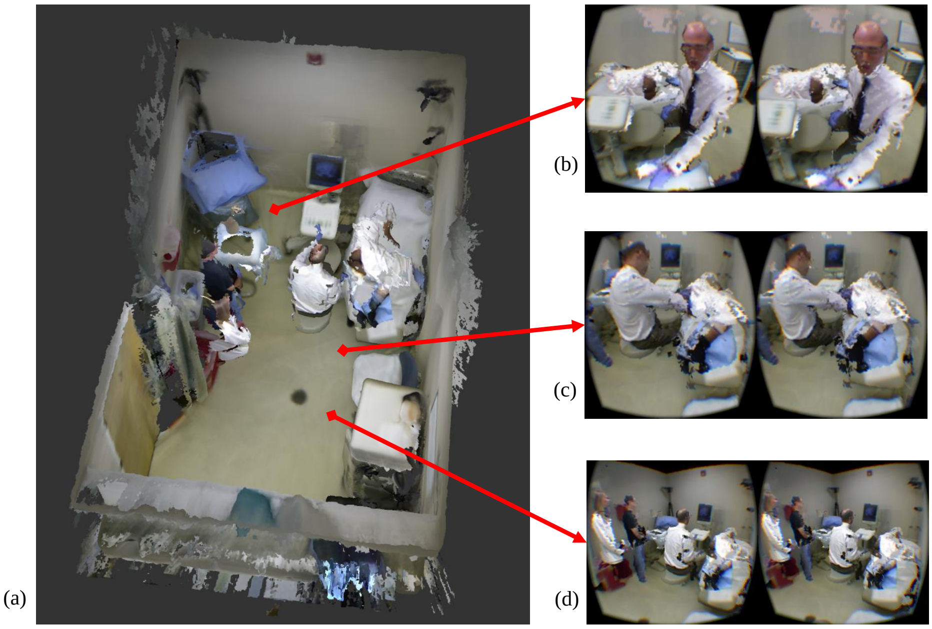


Figure 6. Reconstructed immersive environment. (a): top view with sample user locations (red). (b-d): corresponding views inside HMD. (These are "screen shots" provided by the Oculus SDK, approximations to the images sent to the Oculus screen.)

ground, and three fixed wall-mounted units for dynamic procedure capture [10]. Our multiple Kinect recording setup provided approximately 25 color and depth images per second at 640×480 resolution. The three fixed depth cameras were synchronized manually. The reconstruction and visualization systems used Intel Xeon E5-2630V3 Octa-core 2.4GHz with 64GB memory.

The pre-scanned room shown in Fig. 2a was reconstructed from 401 RGB-depth images captured by a single hand-held Kinect depth camera. The dimensions of the room are approximately $2.5m \times 4.5m \times 3m$ (width, length, and height). In this first experiment 1,841 consecutive multiple-view RGB-depth images were sampled, equivalent to a playback running time of approximately 1.5 minutes. At viewing time, an Oculus Rift DK2 HMD was used and the scene is rendered using the Unity 5 Integration provided by Oculus VR.

Fig. 6 demonstrates the walkaround capability within the reconstructed immersive environment. Fig. 6a shows the reconstructed room (including dynamic objects such as people and instruments), which the user can observe from his own, freely selectable position. Three sample views are depicted in Figs. 6b-d and show the distortion-compensated imagery presented within the Oculus HMD. The user is able to direct his/her gaze at and approach any spatial regions of the scene he/she is interested in, without restrictions.

Fig. 5 illustrates the playback functions while walking around the reconstruction. The user holds a wireless mouse and can click the buttons on the virtual playback controller. This feature makes it easy to find and replay the interesting time snippets.

5. Conclusion and Future Work

In this work, we introduced an immersive learning environment for surgical procedures, created by applying 3D capture and dynamic reconstruction methods to such procedures. The resulting dynamic geometry can be annotated post-reconstruction to enhance educational utility. We hope that such immersively experienced, annotated procedures will be useful for beginning medical students and nurses especially, as it can supplement preparation for their initial patient treatment encounters. For the longer term, we hope that ubiquitous deployment and continuous operation of such acquisition and reconstruction technology will make it possible to re-experience even difficult or unusual cases, helping medical personnel develop skills for interventions that occur infrequently.

In the future, we plan to deploy a larger number of cameras, including higher resolution cameras in order to reduce artifacts caused by occlusion or reconstruction failures. To improve the surface quality of the dynamic scene elements, non-rigid registration methods [18,19,20] can be utilized to continually track and integrate the surfaces of moving objects. We also plan to combine color based multi-view segmentation to improve the surface quality in dynamic scene reconstruction [21]. Finally, we also intend to integrate contributions from cameras and/or depth scanners worn by the attending personnel, which will help improve the most important parts of the reconstructed geometry, since they represent the focus of attention of the medical personnel at the time of the procedure.

Acknowledgments

This research was supported in part by Cisco Systems, by Microsoft Research, by NSF Award II-1405847 (“Seeing the Future: Ubiquitous Computing in EyeGlasses”), and by the BeingThere Centre, a collaboration among ETH Zurich, NTU Singapore, and UNC Chapel Hill, supported by ETH, NTU, UNC, and the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative.

The authors would like to thank Christopher Paterno, for assisting with UNC Medical Institutional Review Board approval procedures and for collaboration on recordings in the clinic, and Jim Mahaney for management of recording equipment for our 3D capture experiments.

References

[1] DEV Parvati, W LeRoy Heinrichs, and Y Patricia, “Clinispace: A multiperson 3d online immersive training environment accessible through a browser,” *Medicine Meets Virtual Reality 18: NextMed*, vol. 163, pp. 173, 2011.

[2] Alexandrova IV, M. Rall, M. Breidt, G. Tullius, U. Kloos, H.H. Blthoff, and B.J. Mohler, “Enhancing medical communication training using motion capture, perspective taking and virtual reality,” *Medicine Meets Virtual Reality (MMVR 19)*, vol. 173, pp. 16, 2012.

[3] Yuan Liu, “Virtual neurosurgical education for image-guided deep brain stimulation neurosurgery,” in *IEEE International Conference on Audio, Language and Image Processing (ICALIP)*, 2014, pp. 623–626.

[4] Lars C Ebert, Tuan T Nguyen, Robert Breitbeck, Marcel Braun, Michael J Thali, and Steffen Ross, “The forensic holodeck: an immersive display for forensic crime scene reconstructions,” *Forensic science, medicine, and pathology*, vol. 10, no. 4, pp. 623–626, 2014.

[5] Andrea Ferracani, Daniele Pezzatini, and Alberto Del Bimbo, “A natural and immersive virtual interface for the surgical safety checklist training,” in *Proceedings of the 2014 ACM International Workshop on Serious Games*, 2014, pp. 27–32.

[6] Qiufeng Lin, Zhoubing Xu, Bo Li, Rebeccah Baucom, Benjamin Poulose, Bennett A Landman, and Robert E Bodenheimer, “Immersive virtual reality for visualization of abdominal ct,” in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2013, vol. 8673, p. 17.

[7] Greg Welch, Dan Russo, Jesse Funaro, Andries van Dam, Adrian Ilie, Kok-Lim Low, Anselmo Lastra, Bruce Cairns, Herman Towles, and Henry Fuchs, “Immersive electronic books for surgical training,” *IEEE MultiMedia*, vol. 12, no. 3, pp. 22–35, 2005.

[8] Gregorij Kurillo, Ruzena Bajcsy, Oliver Kreylos, and Rodny Rodriguez, “Teleimmersive environment for remote medical collaboration,” *Medicine Meets Virtual Reality (MMVR 17)*, vol. 142, pp. 148–150, 2009.

[9] Henry Fuchs, Andrei State, and Jean-Charles Bazin, “Immersive 3d telepresence,” *IEEE Computer*, vol. 47, no. 7, pp. 46–52, 2014.

[10] Mingsong Dou and Henry Fuchs, “Temporally enhanced 3d capture of room-sized dynamic scenes with commodity depth cameras,” in *IEEE Virtual Reality (VR)*, 2014, pp. 39–44.

[11] Aparna Taneja, Luca Ballan, and Marc Pollefeys, “Image based detection of geometric changes in urban environments,” in *IEEE International Conference on Computer Vision (ICCV)*, 2011, pp. 2336–2343.

[12] Ali Osman Ulusoy and Joseph L Mundy, “Image-based 4-d reconstruction using 3-d change detection,” in *European Conference on Computer Vision (ECCV)*, pp. 31–45. Springer, 2014.

[13] 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 *10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011, pp. 127–136.

[14] Mingsong Dou, Li Guan, Jan-Michael Frahm, and Henry Fuchs, “Exploring high-level plane primitives for indoor 3d reconstruction with a hand-held rgb-d camera,” in *Asian Conference on Computer Vision (ACCV) Workshops*, 2012, pp. 94–108.

[15] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, no. 11, pp. 2274–2282, 2012.

[16] William E Lorensen and Harvey E Cline, “Marching cubes: A high resolution 3d surface construction algorithm,” in *ACM SIGGRAPH*, 1987, vol. 21, pp. 163–169.

[17] Brian Curless and Marc Levoy, “A volumetric method for building complex models from range images,” in *ACM SIGGRAPH*, 1996, pp. 303–312.

[18] Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, and Shahram Izadi, “3d scanning deformable objects with a single rgbd sensor,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 493–501.

[19] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, and Christian Theobalt, “Real-time non-rigid reconstruction using an rgb-d camera,” *ACM Transactions on Graphics (TOG)*, vol. 4, 2014.

[20] Richard A Newcombe, Dieter Fox, and Steven M Seitz, “Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 343–352.

[21] Abdelaziz Djelouah, Jean-Sébastien Franco, Edmond Boyer, Francois Le Clerc, and Pablo Perez, “Multi-view object segmentation in space and time,” in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 2640–2647.