A Portable Immersive Surgery Training System Using RGB-D Sensors

Xinqing GUO [a], Luis D. LOPEZ [a], Zhan YU [a], Karl V. STEINER [a], Kenneth E. BARNER [a], Thomas L. BAUER [b] and Jingyi YU [a]

[a] *University of Delaware*
[b] *Christiana Care Health Services*
Newark, Delaware

Abstract. Surgical training plays an important role in assisting residents to develop critical skills. Providing effective surgical training, however, remains as a challenging task. Existing videotaped training instructions can only show imagery from a fixed viewpoint that lacks both depth perception and interactivity. We present a new portable immersive surgical training system that is capable of acquiring and displaying high fidelity 3D reconstructions of actual surgical procedures. Our solution utilizes a set of Microsoft Kinect sensors to simultaneously recover the participants, the surgical environment, and the surgical scene itself. We then develop a space-time navigator to allow the trainees to witness and explore a prior procedure as if they were there. Preliminary feedback from residents shows that our system is much more effective than conventional videotaped system.

Keywords. RGB-D Sensor, Microsoft Kinect, Immersive Surgery Training, 3D Reconstruction, Stereoscopic Display

Introduction

In the U.S., surgeons require longer education and training than other specialists: only after four years of medical school and a minimum of five years of extensive training will they qualify. The satisfaction of surgical residents with their training program determines its output. The task of providing effective surgical training and re-training, however, is inherently challenging: the number of high quality educators is rather limited and both instructors and trainees are over-constrained by time. The problem is further deteriorating as new surgical procedures are becoming increasingly complex and often require using new devices and protocols.

In traditional surgical training, videotaped instruction has long served as a workhorse for teaching surgical procedures. However, they are marginally effective: videotapes only provide 2D imagery that lacks depth perception and the trainee cannot freely change viewpoints as the inputs are captured from a fixed location. To address these issues, the pioneering work of 3D telepresence [1,2,3,4] aims to emulate remote medical procedures. At its core are acquisition, reconstruction, and display of the complete 3D geometry in room-sized surgical environments. Most existing approaches [5,2,6,7,8], e.g., from Fuchs's group at UNC, Bajcsy's group at Penn, Kanade's group at CMU, and Gross's group at ETH, have pioneered the use of a "sea of cameras" around a room. Their sem-

inal work has led to great advances in telemedicine and provides useful insights on the system design and processing algorithms.

However, there is not a single educational environment that comes close to replacing the traditional apprenticeship environment of the Operating Room (OR) for two main reasons. On the system front, it is literally impractical to mount "a sea of cameras" within an OR. Most existing multi-camera systems (including the immersive solutions mentioned above) require using multiple workstations just for data transmission and storage. The system infrastructure, such as camera mountings, interconnects, and workstations, is bulky, making them unsuitable for on-site tasks. On the reconstruction front, recovering 3D scene geometry from images is still one of the open problems in computer vision [9,10]. To make the problem tractable, many existing algorithms tend to make simplified assumptions about scenes, such as Lambertian surface and distant light sources. However, in surgical environments, we simply cannot assume these factors. For example, specular highlights and changing lighting are the norm in surgical environments (with body fluid, metal instruments, head-mounted lights, etc.), easily causing classical computer vision algorithms, such as binocular stereo or shape-from-shading to break down.

In this paper, we present a new immersive surgical training system. Our proposed solution resolves both the system and reconstruction problems by leveraging emerging 3D imaging technologies and multi-modal fusion algorithms. Instead of using a large number of cameras, we use a small number ($2 \sim 4$) of 3D sensors, namely Microsoft Kinect. These sensors are uniformly controlled by a single workstation and their range and imagery data are fused via a companion computer vision algorithm for robustly recovering the 3D surgical scene. We further develop a user interface to allow the trainees to navigate the 3D environment in both space and time.

Our preliminary experiments, conducted at the Virtual Education and Simulation Technology (VEST) Center at Christiana Care Health System (CCHS), show that our system can effectively capture and reconstruct 3D surgical procedures performed by an expert. These three-dimensional recordings can be presented in a virtual operation theater in which medical students can perceive solid stereoscopic views without glasses (e.g. on an autostereo display) or with special glasses on a commercial 3D TV, as if they were present in the room.

1. Methods and Materials

Figure 1 shows our proposed immersive surgical training system that can automatically recover 3D surgical scenes. Our system consists of three major components. The first component, Image Acquisition, captures images and depth data using a set of Microsoft Kinect cameras and recovers the camera calibration matrix for each view. Next, the Data Fusion and 3D stereoscopic rendering module combines the image and depth data to generate a 3D point cloud from each view and utilizes the camera calibration parameters to fuse individual data into a global 3D point cloud, which is subsequently rendered as a 3D stereoscopic view of the scene. Finally, the Data Navigation module allows users to dynamically visualize the surgical event from new perspectives at arbitrary time instances in real time.

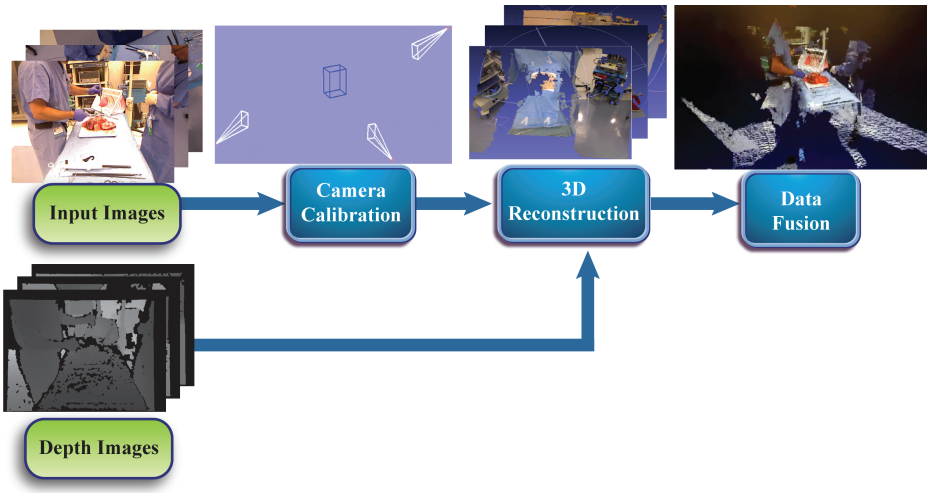


Figure 1. Our proposed pipeline for reconstructing and visualizing 3D surgical environments.

1.1. Image Acquisition and Camera Pose Recovery

Figure 2 shows our image acquisition system that uses a set of three Microsoft Kinect sensors. Each Kinect sensor consists of an infrared projector, a RGB camera with a resolution of 640×480 pixels, and an infrared sensor. Also, a calibration pattern is used to determine point correspondence to automatically recover the camera calibration parameters. To get access to both depth and RGB image streams, we develop our data fetching module based upon the open source nestk library [11].

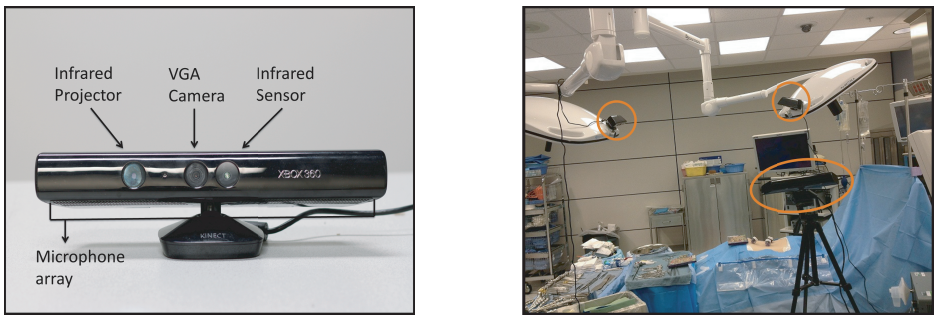
In our experiments we strategically mount the Kinect sensors around the operating table to cover both the organs and surgeons during the surgical procedure. A computer with an i7-3930k processor is used to communicate with the Kinect sensors through USB interfaces. During acquisition, both depth and RGB images are captured at a rate of 15 frames per second for all Kinect sensors.

Similar to previous approaches [2,3,12,13,14], our method requires obtaining the camera calibration matrix for each view. In our solution, the OR have very similar colors without textures and the occlusion patterns vary significantly across views due to sparse sampling, making it challenging to robustly compute the point correspondence across views. To resolve this issue, we manually identify point correspondences between the corners of the calibration pattern in the acquisition system. We then recover the camera calibration parameters for each view, using the approach described in [15].

1.2. Data Fusion and 3D Stereoscopic Rendering

Next, we perform a two pass rendering approach that first recovers a point cloud for each viewpoint, and then fuses the individual results to generate a dense set of 3D points that faithfully reconstruct the 3D scene.

Given a depth-RGB image pair, our solution first traces a ray for each pixel in the image and utilizes the depth data to find the corresponding 3D coordinates. Specifically,



(a) (b)

Figure 2. (a) Microsoft Kinect has a microphone array, an infrared projector, an infrared sensor and a VGA camera. (b) Acquisition system consists of a set of three Microsoft Kinect cameras.

for each pixel in the input image we trace a ray originating at the center of projection $\mathbf{C}$ toward the image plane. Let $\bar{\mathbf{r}}$ denote a ray originating from $\mathbf{C}$ toward pixel (u,v) in the image plane. The trajectory of the ray can be described as

$$\bar{\mathbf{r}} = \mathbf{C} + \lambda \bar{\mathbf{d}} \tag{1}$$

where $\bar{\mathbf{d}}$ is the direction vector. In camera coordinate system, the direction vector $\bar{\mathbf{d}}$ can be written in terms of camera image plane axis $\bar{\mathbf{d}}_\mathbf{x}$, $\bar{\mathbf{d}}_\mathbf{y}$ and the optical axis $\bar{\mathbf{d}}_\mathbf{z}$ as:

$$\bar{\mathbf{d}} = u\bar{\mathbf{d}}_\mathbf{x} + v\bar{\mathbf{d}}_\mathbf{y} + f\bar{\mathbf{d}}_\mathbf{z} \tag{2}$$

Here (u,v) is the pixel coordinate in the image plane and f is the focal length of the camera. Therefore, the original equation can be described as

$$\bar{\mathbf{r}} = \mathbf{C} + \lambda (u\bar{\mathbf{d}}_\mathbf{x} + v\bar{\mathbf{d}}_\mathbf{y} + f\bar{\mathbf{d}}_\mathbf{z}) \tag{3}$$

Notice that the ray intersects the image plane when $\lambda = 1$. Since the depth image contains a measure of depth along the optical axis, we can conveniently determine λ for each pixel. Thus, for each Kinect sensor we can compute a 3D textured point from each input pixel in the 2D image.

Next, we use the camera calibration parameters to transform the point cloud of each Kinect sensor from local coordinate into a global coordinate system. Then we fuse multiple point clouds into one global point cloud representation. Notice that Kinect is designed as a stand-alone solution. While a single Kinect sensor delivers quite robust depth maps, simultaneously running multiple sensors may lead to deteriorated results.

With the generated point cloud, we set out to render a 3D stereoscopic view of the scene. Traditional 3D rendering generates a single perspective view by synthesizing a pinhole camera image in the scene. We extend this approach by simultaneously setting two cameras in the scene with a user specified baseline. In a single frame, two cameras capture two views of the point cloud and render them with red-cyan anaglyph. We also

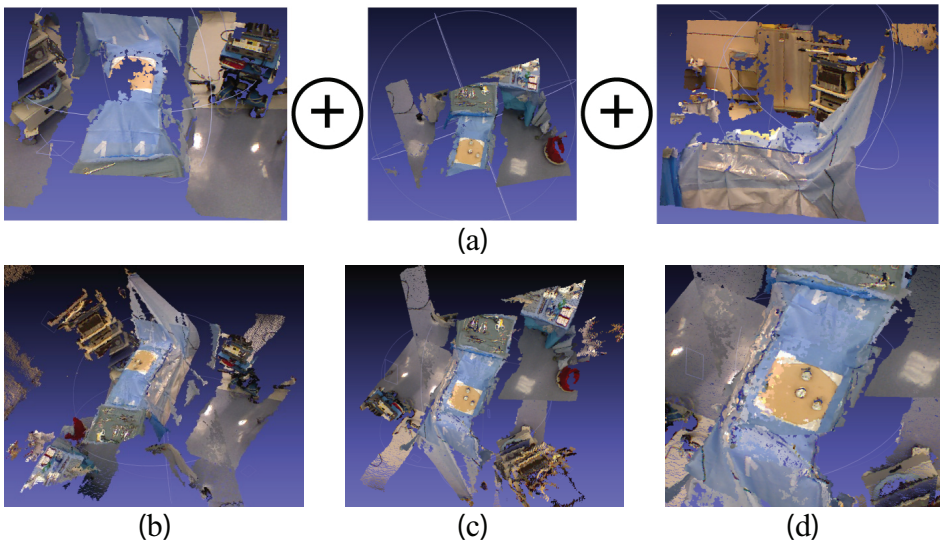


Figure 3. (a) Combining point clouds acquired from each Kinect sensor. (b) A global point cloud representation. (c) Change of viewpoint. (d) A close-up view.

utilized the NVIDIA 3D API to render two regular color images and synchronize with the NVIDIA 3D glasses to deliver a better user experience. Both passes are mapped onto Microsoft Direct3D graphics framework. Using the state of the art NVIDIA GeForce GTX 580 graphics card, we can render stereoscopic views at over 1000 fps with a resolution of 1280×1024.

1.3. Data Navigation

To better help trainees to understand a surgical procedure, we have developed a space-time visualization system to display the acquired data. The system includes an interface, which allows users to pick a specific time frame in a surgical procedure, pause or replay that time frame and dynamically change viewpoints and have close-up views as if they were there. Our new navigation system thus allows the trainees to review a surgical procedure without any space or time constraint, as shown in the videos from our project website [16].

2. Results and Discussions

Our ultimate goal is to bring together researchers and clinical trainees to evaluate the proposed system. To that end, we have worked closely with the VEST Center at CCHS, which supports the entire Christiana Care Community (physicians, nurses, allied health professionals, residents, students and regional health services). The VEST Center includes adult and pediatric high-fidelity human patient simulators, a working laparoscopy station with simulated tissues, an endoscopy/bronchoscopy simulator, 3D visualization software and display and numerous task trainers to meet departmental needs. In addition,

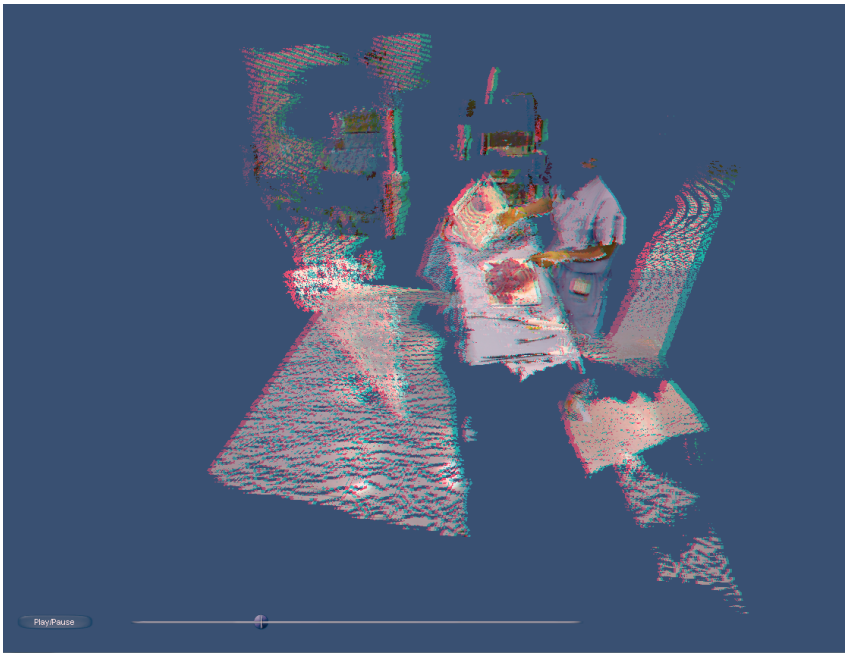


Figure 4. 3D stereoscopic view using red-cyan anaglygh.

the VEST center has two operative theaters approved for tissue block surgery, fully fitted with all instrumentation and equipment for surgical procedures.

We used our system to capture a cholecystectomy (gallbladder surgery) on animal tissue blocks conducted by highly trained surgeons at the VEST Center. To cover as many details as possible on the operating table, we used three Kinect sensors facing the table. For training purposes, the surgery took half an hour and we were able to capture five video clips. Figure 3(a) shows three point clouds acquired from the three Kinect sensors. Figure 3(b) shows the global point cloud representation by combining three point clouds. As shown in Figure 3(c) and 3(d), one can change viewpoints and zoom in and out using our system. Figure 4 shows the 3D stereoscopic view using red-cyan anaglyph. Initial Feedback from the residents shows that our system is much more effective than the conventional videotaped system. These results along with additional videos can be found at [16].

3. Conclusions and Future Work

We have developed a new immersive surgery training system by coupling emerging 3D imaging technologies with advanced computer vision and graphics techniques. Specifically, we use the Microsoft Kinect platform, an inexpensive commercial 3D camera, as the main acquisition device and develop a class of multi-view 3D fusion techniques to faithfully reconstruct the surgical procedure. We have conducted preliminary tests of the system fidelity for cholecystectomy (gallbladder surgery) training and have developed a space-time visualization system to display the acquired data. Furthermore, we integrate

our system with 3D stereoscopic displays to enhance the user experience. For the next stage, We will explore possible integrations with the Visible Human [17] and the Digital Anatomist [18] projects.

4. Acknowledgments

This project is supported by the Delaware INBRE under a grant from NIGMS (8P20 GM103446) at NIH.

References

[1] H. Fuchs and U. Neumann. A vision of telepresence for medical consultation and other applications. In *Proceedings of the Sixth International Symposium on Robotics Research*, pages 565–571, 1993.

[2] Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. The office of the future: a unified approach to image-based modeling and spatially immersive displays. ACM SIGGRAPH, pages 179–188, 1998.

[3] L.-Q. Xu, B. Lei, and E. Hendriks. Computer vision for a 3-d visualisation and telepresence collaborative working environment. *BT Technology Journal*, 20(1):64–74, January 2002.

[4] E. Trucco, K Plakas, Nicole Brandenburg, Peter Kauff, Michael Karl, and Oliver Schreer. Real-time disparity maps for immersive 3-d teleconferencing by hybrid recursive matching and census transform. In *ICCV, Proceeding of Workshop on Video Registration*, 2001.

[5] Henry Fuchs, Gary Bishop, Kevin Arthur, Leonard McMillan, Henry Fuchs Gary Bishop, Ruzena Bajcsy, Sang Wook Lee, Hany Farid, and Takeo Kanade. Virtual space teleconferencing using a sea of cameras. In *Proc. First International Conference on Medical Robotics and Computer Assisted Surgery*, pages 161–167, 1994.

[6] Oliver G. Staadt, Markus H. Gross, Andreas Kunz, and Markus Meier. The blue-c (poster session): integrating real humans into a networked immersive environment. In *Proceedings of the third international conference on Collaborative virtual environments*, CVE '00, pages 201–202, 2000.

[7] Kok lim Low, Adrian Ilie, Greg Welch, and Anselmo Lastra. Combining head-mounted and projector-based displays for surgical training. In *in Proceedings of IEEE Virtual Reality 2003. Los*, pages 110–117, 2003.

[8] Greg Welch, Andrei State, Adrian Ilie, Kok-Lim Low, Anselmo Lastra, Bruce Cairns, Herman Towles, Henry Fuchs, Ruigang Yang, Sascha Becker, Dan Russo, Jesse Funaro, and Andries van Dam. Immersive electronic books for surgical training. *IEEE MultiMedia*, 12(3):22–35, July 2005.

[9] Motilal Agrawal and Larry S. Davis. A probabilistic framework for surface reconstruction from multiple images. In *CVPR (2)*, pages 470–476, 2001.

[10] P. J. Narayanan, Peter W. Rander, and Takeo Kanade. Constructing virtual worlds using dense stereo. In *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, pages 3–10, 1998.

[11] Nestk library. https://github.com/nburrus/nestk.

[12] Yuanjie Zheng, Chandra Kambhamettu, Jingyi Yu, Thomas Bauer, and Karl Steiner. Fuzzymatte: A computationally efficient scheme for interactive matting. In *CVPR*, 2008.

[13] Yuanjie Zheng, Jingyi Yu, Chandra Kambhamettu, Sarah Englander, Mitchell D. Schnall, and Dinggang Shen. De-enhancing the dynamic contrast-enhanced breast mri for robust registration. In *Proceedings of the 10th international conference on Medical image computing and computer-assisted intervention - Volume Part I*, MICCAI'07, pages 933–941, 2007.

[14] Yuanjie Zheng, Karl Steiner, Thomas Bauer, Jingyi Yu, Dinggang Shen, and Chandra Kambhamettu. Lung nodule growth analysis from 3d ct data with a coupled segmentation and registration framework. In *ICCV*, pages 1–8, 2007.

[15] Q.-T. Luong and O. D. Faugeras. Self-calibration of a moving camera from pointcorrespondences and fundamental matrices. *Int. J. Comput. Vision*, 22(3):261–289, March 1997.

[16] Immersive Surgery Training System. http://www.eecis.udel.edu/∼xinqing/inbre/.

[17] The Visible Human Project. http://www.nlm.nih.gov/research/visible.

[18] Digital Anatomist Project. http://sig.biostr.washington.edu/projects/da/index.html.