A non-contact interactive stereo display system for exploring human anatomy

Ziteng Liua, Wenpeng Gaoa, Yu Sunb, Yixian Sub, Jiahua Zhub, Luming Xua and Yili Fuab

aSchool of Life Science and Technology, Harbin Institute of Technology, Harbin, China; bState Key Laboratory of Robotics and System, Harbin Institute of Technology, Harbin, China

ABSTRACT

Stereoscopic display based on Virtual Reality (VR) can facilitate clinicians observing 3D anatomical models with the depth cue which lets them understand the spatial relationship between different anatomical structures intuitively. However, there are few input devices available in the sterile field of the operating room for controlling 3D anatomical models. This paper presents a cost-effective VR application for stereo display of 3D anatomical models with non-contact interaction. The system is integrated with hand gesture interaction and voice interaction to achieve non-contact interaction. Hand gesture interaction is based on Leap Motion. Voice interaction is implemented based on Bing Speech for English language and Aitalk for Chinese language. A local database is designed to record the anatomical terminologies organized in a tree structure, and provided to the speech recognition engine for querying these uncommon words. Ten participants were asked to practice the proposed system and compare it with the common interactive manners. The results show that our system is more efficient than the common interactive manner and prove the feasibility and practicability of the proposed system used in the sterile field.

KEYWORDS

Virtual reality; stereo display; gesture interaction; voice interaction; 3D anatomical model

Introduction

Recently, medical imaging, such as computer tomography (CT) and magnetic resonance imaging (MRI), plays an important role in modern medicine, especially in the diagnosis and therapy for its high resolution in differentiating the human anatomies in vivo without invasion. However, films as the prevalent medical image display modes are not intuitive for understanding, which impose large mental burden on the surgeons to imagine the shape of the anatomy and its spatial relationship with its neighboring anatomies.

Modern trend in medical image display focuses on VR which is used to create and show a 3D model of the organ from medical images and provides the clinicians with more detail information of the interesting organ than 2D films. In this way, clinicians can observe 3D anatomical models in an immersive way and understand the spatial position relationship between the anatomical structures intuitively. Ioannou et al. [1] proposed a VR-based simulator to help the dental and oral surgeons to train and practice more variously and professionally than before. Xia et al. [2] introduced a VR workbench to assist surgeons making surgical planning for orthognathic surgery. Hong et al. [3] proposed a VR and augmented reality surgical navigation solution and provided the surgeons with more real-time surgical information. Moreover, VR can also facilitate medical education.

To better understand the human anatomies, common input devices, such as the mouse, keyboard or buttons, are usually used to control the VR environment. However, in the operating room where the environment is sterile, it is not convenient for the surgeons to maneuver these devices. Some non-contact interaction methods have come to being, for example, gesture interaction and voice interaction. Ruppert et al. [4] proposed a non-contact interactive VR display system based on Kinect (Microsoft Corp, USA), which allows surgeons to interact with the system in an efficient and natural manner to avoid being forced to divert their attention during the operation. Schuller et al. [5] proposed a voice control interface for minimal invasive surgery and demonstrated its practicability. Certainly, each of the approaches has its own advantages and disadvantages. Gesture interactive
manner has the advantage over speech manner in fine-tuning, such as rotating the model in a proper view angle according to the visual feedback, whereas the voice interactive manner is good at logical operation, for example, switching operation mode. It will make VR more powerful in medical image display if these two non-contact human-computer interactive modes could be integrated into the display system.

To the best of our knowledge, there is no report on the application of VR combined with gesture and voice interaction in the medical application. We proposed a VR system for 3D anatomical model demonstration integrated with non-contact interaction methods, including gesture interaction and voice interaction. Experimental results show that the integration of the gesture recognition and speech recognition to the VR system is an efficient and convenient way that can help clinicians to observe the virtual anatomical models when they are not convenient to use the common input device to control them in the operating room.

System implementation

The prototype of the proposed system is composed of a head-mounted display (HMD), a hand tracking device and a computer, as shown in Figure 1. Oculus Rift DK2 (Facebook Inc., USA) is used as the HMD with a high-resolution (960 × 1080 per eye) low-persistence OLED display, a refresh rate of 90 Hz, a field of view (110°), an internal IMU and an external optical sensor for head orientation and position tracking respectively. It supports 3D stereoscopic display for VR demonstration to let the user immerse in the virtual scene completely. The hand-tracking device is a Leap Motion controller (Leap Motion Inc., USA), which consists of two monochromatic IR cameras for hand gesture recognition. The computer is with Core i7-6700, Nvidia GeForce GTX 1060 6GB and 16GB RAM, and a microphone for audio signal acquisition. It is in charge of VR rendering, hand gesture recognition, speech recognition and user interaction. As shown in Figure 1, this system can provide the doctors with 3D anatomical models and allow the doctors to interact with the system efficiently through hand gesture and voice command.

The proposed system consists of four modules, namely, local database, virtual scene rendering module (VSR), gesture interaction module (GIM) and voice interaction module (VIM). The pipeline of the proposed system is shown in Figure 2.

Local database

The local database is a relation database designed to record the anatomical terminologies in a tree structure, the visibility status and the file path of each model. The hierarchical nature of the anatomical terminologies is often organized in a tree structure. Spinal terminologies are taken as example to explain the tree structure and the design of the database. A normal adult’s spine composed of 33 bones is divided into five sections, i.e., cervical region (7 vertebrae), thoracic region (12 vertebrae), lumbar region (5 vertebrae), sacrum (5 fused bones) and coccyx (4 fused bones). The relationship between each bones is represented in a tree structure shown in Figure 3. The related table of the database consists of seven items, i.e., ID, Name, Sub-structure ID, Parent structure ID, Visible, File path and Description (see Figure 3). Each anatomies in the database has a unique ID. Name records the terminology of the anatomy which provides a glossary for speech recognition engine. Sub-structure ID and Parent structure ID record the IDs of its sub-structures and the structure it belongs to, respectively. This guarantees the anatomy composed of multi sub-structures can be shown as an entity in the scene by recursive retrieving its all sub-structures. Visible is consulted by VSR to check whether the 3D model of the anatomy is displayed in the scene or not. VIM sets value to this item to let it be shown or not. File path helps the system to locate the file of 3D model data storage in the computer. Description is the description of the anatomy. The proposed system is initialized from the database to load all anatomical models into the scene.
Virtual scene rendering module

This module is responsible for creating and updating the virtual scene containing 3D anatomical models and transferring it to the Oculus Rift DK2 screen. It is implemented using OpenGL render pipeline. To correctly simulate a 3D environment and display it to the user, the display system uses the stereo parallax to provide the user with the depth cue. To preserve the depth clue from shading, 3D anatomical models in the scene are rendered using Blinn-Phong shading model [6] with the support of OpenGL Shader Language and two virtual cameras are set with their principle axis parallel and the baseline tunable to render a pair of images of the scene.

VSR updates the scene accordingly when the model transformation matrix is changed based on the users’ hand gesture. The model transformation matrix $M_m$ is computed by:

$$M_m = M_{sh} \cdot M_s \cdot M_r$$  \hspace{1cm} (1)

where $M_{sh}$ is a model shift matrix, $M_s$ is a model scale matrix, $M_r$ is a model rotation matrix. The mapping from the users’ hand gesture to the matrix is introduced in the gesture interaction module. The scene is also updated when the cameras’ pose changes. The cameras’ orientation and position are in consistent with those of the HMD tracked by the internal IMU and external optical sensor, respectively. After the VSR updates and renders the virtual scene to FBOs, the FBOs are transferred to the Oculus Texture Chain by calling Oculus SDK and the HMD serves a pair of parallax image to wearer. A virtual spine model and a virtual brain model are shown in Figure 4 as the examples.

Moreover, four modes are defined to make the interaction logic simple and clear: display mode (DM), scale mode (SM), shift mode (ShM) and rotation mode (RM). In the DM, the anatomical models in the scene are kept static, meaning that the transformation matrix $M_m$ in Eq. (1) is locked. In the SM, the anatomical models can only be zoomed in or out, meaning that $M_{sh}$ and $M_s$ in Eq. (1) are locked. Similarly, in the ShM (or RM), the anatomical models can only be shifted (or rotated). To avoid coupling operations, only one mode executes in VSR at any time. Nonetheless, switching among these modes is allowed.
Gesture interaction module

GIM is used to map the user’s hand gesture to the transformation matrix $M_m$ in Eq. (1) and triggers VSR to update the scene. Leap Motion is the core component of the GIM. It can recognize the hand gesture and track the hand position via stereo infrared cameras image. Two hand gestures are tracked in the GIM: grabbing and moving, as shown in Figure 5. Only the hand closest to the cameras is taken into account if GIM detects more than one hand. When the VSR is not in the DM, the GIM recalculates and updates $M_m$ by recording and processing the initial hand position $H_i$ and current hand position $H_c$ of grabbing gesture during the period of GG’s appearance.

In the RM, we adopt a trackball to handle rotation and use a nonlinear mapping algorithm that maps the hand position from Leap Motion coordinate system to the trackball coordinate system through a piecewise quadratic function. After mapping $H_i$ and $H_c$ to the trackball, the GIM computes $M_r$. In the SM, a scaling coefficient is calculated by the magnitude and direction of the displacement vector $H$ from $H_i$ to $H_c$. Then, the GIM computes $M_s$ by multiplying scaling coefficient by an identity matrix. In the ShM, the GIM translates $H$ into $M_{sh}$ and multiplied by an optional weighting coefficient to adjust shift sensitivity.

Voice interaction module

VIM allows user to interact with the system by using voice command. In the VIM, Bing Speech (Microsoft Corp, USA) and Aitalk (IFLYTEK CO., CHINA) are adopted for English and Chinese speech recognition, respectively. Bing Speech is an online speech recognition engine, meaning that the recognition results should be obtained from a server through POST (HTTP) request method. Whereas, Aitalk provides offline service. Not matter online or offline, these two engines cannot return accurate results, because most commands in the medical scenario contain many rarely used anatomical terminologies. Therefore, we created a Bing Language Model for Bing Speech and a Backus-Naur form semantic document for Aitalk so that the recognition engines can directly query the anatomical terminologies in the LRD. Then, an anatomical structure (such as spine) with several sub structures can be shown as an entity when it is expected to be shown in the voice command.

VIM only accepts two kinds of commands: switch system mode command (SSMC) and select model command (SMC). The former is to trigger the VSR to switch the mode. For example, when user says “scale mode”, then VSR switches to the scale mode in which the user can zoom out and in the scene. The latter is to notify VSR showing the model of the desired anatomy. The format of SMC is “show + (name of the anatomy)”. If the anatomy contains sub structures, these sub structures will also be shown in the scene.

Experiments and results

To evaluate the performance of the proposed system, we performed several user studies with a spine anatomy model that consists of 27 anatomic structures. 10 participants were enrolled to attend the test, including 2 doctors, 3 students with medicine background and 5 students with no medicine background. Firstly, all participants were asked to perform a serious of pre-designed operations both in our system and in 3D Slicer, which is a popular software package for image analysis.

Figure 4. The scene with the user’s hand model indicating the hand gesture and a spine model (left) and a brain model (right).

Figure 5. Gestures and windows coordinate mapping in GIM. Left: Grabbing gesture. Right: Moving gesture.
and scientific visualization, to evaluate our system. Then, all participants answered questionnaires to carry out a qualitative analysis of system’s user experience.

**Gesture efficiency evaluation**

All participations were asked to perform the following operations 20 times: (1) Rotating the model to its back side. (2) Zooming out the model from half of the window size to fulfill the window. (3) Moving the model from center to top. In this task, moving the model to the boundary means the model is just entirely out of the view. All of these tasks were processed in the initial state of the system and of 3D Slicer. To speed up this task, we enabled keyboard shortcut for SSMC so that participants did not need to click their mouse.

It takes $0.75 \pm 0.13$ s vs. $1.39 \pm 0.37$ s, $0.83 \pm 0.22$ s vs. $1.12 \pm 0.20$ s, and $0.71 \pm 0.10$ s vs. $1.48 \pm 0.30$ s to perform rotation, shift and scale operation with the proposed system vs. 3D Slicer. The time consumption of two methods was significantly different ($p < .001$, t-test). Figure 6 summarized the consumed time. 22 outliers (3.7%) and 20 outliers (3.3%) were detected in the task of our system and 3D Slicer, respectively. The interquartile ranges (IQR) of our system were ranged from 0.68 s to 0.78 s for rotation, from 0.73 s to 0.88 s for shift and from 0.63 s to 0.77 s for scale. By contrast, 3D Slicer’s IQR were ranged from 1.13 s to 1.60 s for rotation, from 1.00 s to 1.21 s for shift and from 1.28 s to 1.62 s for scale.

**Voice practicability**

In this task, participants were asked to perform a list of commands that contained all the possible voice command for three times. The command list was built by random sampling all possible SSMCs once and four types of SMCs three times. We recorded the number of attempts and response time to measure its success rate and latency. The response time was defined as time span between participants finished speaking and system successfully responded to the command. Both of the time spans and the success rate were taken into account to evaluate the practicability.

After completing the tasks, 3680 trials were gleaned with a command list, which contained 184 voice commands performed by 10 participants with Bing Speech and Aitalk, respectively. The recognition time (mean±standard deviation) of Bing Speech and Aitalk were $0.82 \pm 0.24$ s and $0.21 \pm 0.08$ s, respectively. 57 trails (3.1%) were performed twice and 21 trails (1.1%) at least three times by using Bing Speech. By contrast, 45 trails (2.4%) were performed twice and 17 trails (0.9%) at least three times by using Aitalk. Two sample $t$-test indicated that there was significant difference between two speech recognition engines in term of the success rate ($p < .001$).

**User experience evaluation**

After the system practice, all participants were asked to answer the questionnaires with following score items: (1) **Display Effect**. Participants graded this item according to the model display quality such as the completeness of model detail and discrimination of different structures. (2) **Interaction Experience**. Several factors are taken into account, such as interaction fluency, interaction convenience and user friendliness. (3) **Practicability in Operating Room**. Only doctors and medical students were asked to grade this item because of professional requirements. The score represented the level of practicability that applies this system into operating room. All items were rated in a range from 0 to 10.

The results were shown in Table 1. The proposed system won positive comments. According to the score of the Display Effect, all participants believed that the VR with the Oculus Rift DK2 are superior to 2D display in the illustration of 3D anatomical
models. As for Interaction Experience, participants thought our system is better. In term of Practicability in Operating Room, they gave our system a higher grade.

Discussion

The results of gesture efficiency evaluation show that the presented system is more efficient than 3D Slicer in controlling 3D models display. Particularly, voice interaction is straightforward and more efficient in logical operation than common input devices. Moreover, our system’s IQRs have a smaller variance than 3D Slicer. These results suggest that gesture interaction as a natural interaction manner provides better hand-eye coordination than those controlled by mouse in 3D Slicer.

In Voice Practicability task, the speed of Aitalk is faster than that of Bing Speech due to different working manners of two engines. This is contributed to Aitalk’s offline speech recognition engine without the time waste in the network. Nevertheless, the speed of Bing Speech is still satisfactory. Both of two engines achieve a lower error rate with respect to anatomical terminologies, owing to the application of a local database of the anatomical terminologies, which guarantees that the proposed system can respond to the user’s voice command fast and accurately.

Based on the user experience evaluation task, it can be concluded that the VR display technology used in this system can provide the users with less cognitive burden to help the users observe 3D anatomical models in an intuitive manner. In addition, non-contact interaction manners also facilitate the surgeons controlling the model in the sterile field in the operating room.

Non-contact interaction techniques show perspective in the application of VR. However, it still takes a short time for users to learn. The speed, accuracy and robustness of these techniques should also be improved to achieve better user experience.

Conclusion

We combined VR technology with speech interactive method and gesture interactive method which are gradually mature in the practical application, and implemented such a non-contact interactive stereo display system for 3D anatomical models. Our system has been tested in several tasks and the results proved that it had a more efficient non-contact interactive mode and an impressive stereo display mode compared with the traditional display system. Our system achieves a convenient, intuitive and non-contact interactive stereo display of virtual anatomy model and facilitates the study and research for doctors and medical students. In the future, we will focus on improving the user experience and extending the application scope.

Disclosure statement

No potential conflicts of interest were disclosed.

Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 81171304, 81201150 and 81500924), China Postdoctoral Science Foundation (grant numbers 2012MS20741 and 2014T70359), Self-Planned Task of State Key Laboratory of Robotics and System (HIT) [grant number SKLRS201407B].

References

[1] Ioannou I, Kazmierczak E, Stern L, et al. Comparison of oral surgery task performance in a virtual reality surgical simulator and an animal model using objective measures[C]. Int Conf IEEE Eng Med Biol Soc. 2015:5114–5117.
[2] Xia J, Ip HH, Samman N, et al. Three-dimensional virtual-reality surgical planning and soft-tissue prediction for orthognathic surgery.[J]. IEEE T Inf Technol B. 2001; 5:97–107.
[3] Pandya A, Auner G. Simultaneous augmented and virtual reality for surgical navigation[C]// Fuzzy Information Processing Society, 2005. Nafips 2005 Meeting of the North American. IEEE. 2005;429–435.
[4] Ruppert GCS, Reis LO, Amorim PHJ, et al. Touchless gesture user interface for interactive image visualization in urological surgery[J]. World J Urol. 2012;187: 687–691.
[5] Schuller B, Can S, Scheuermann C, et al. Robust speech recognition for -robot interaction in minimal invasive surgery[C]// Russian-bavarian conference on bio-medical engineering, Rbc-Bme. 2008:197–201.
[6] Westover L. Footprint evaluation for volume rendering[C]// Conference on Computer Graphics & Interactive Techniques. ACM. 1990; 24:367–376.