Intelligent 3D Online Virtual Conferencing System with Natural Human-Computer Interaction

Yu-Chi Su, Chia-Jeng Tsai, Keng-Yen Huang, and Liang-Gee Chen

Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan
steffi@video.ee.ntu.edu.tw

***Abstract*—In this paper, we propose an intelligent 3D online virtual conferencing system with a natural human-computer interactive mechanism. With 4 degree-of-freedom of gesture commands as the human-computer interface of the system, participants can change their positions and orientations in a virtual space to interact with others and observe the environment. A 3D tracking-based gesture recognition method with high accuracy and a 3D scene construction technique are adopted in the proposed system. With the real 3D interactive experience provided, the system enables people around the world to communicate and cooperate with each other easier and more efficient than ever before.**

Keywords- virtual conferencing; human-computer interface; gesture recognition;

I. INTRODUCTION

With the phenomenon of globalization, it is more and more difficult to gather your coworkers around the world physically together in a conference room. Thus, how to construct an efficient method to communicate with others anytime and anywhere becomes an attractive issue. Several internet technologies have been proposed to provide people some alternative ways to have discussion without getting face-to-face. Social network systems like Facebook or Twitter can connect individual user to the social network by sharing personal photos or videos immediately. Instant messaging is another widely used channel for people to chat with friends without spatial limitation. However, the above methods lack in supporting audio and video communication for multiple users joining together.

Virtual conferencing is more and more popular in hosting a meeting with participants spreading around the world. Existing commercial virtual conferencing systems [1] supports multi-party interaction by showing the face of every participant and their conversation dialogs but lack real interactive experiences in a 3D scene. In the real life, there are different scenarios occurring in a conference. As shown in Fig. 1, one participant can give a presentation, talk to his neighbors, or sit on a special position because of his unique identity. In other words, with different actions in various scenarios, participants should see different angles of view, which helps him to observe others and provide real 3D interactive experience within the virtual space. Considering the genuine circumstances, two previous works [2][3] proposes mixed-reality conferencing systems with head tracking mechanism and the later even combines the tracking method with SpaceMouse. However, head movements are not natural interactive commands for users when they sit in front of computers with staring at displays.

Motivated by this, we propose an intelligent 3D online virtual conferencing system with a natural human-computer interactive mechanism. Participants sitting in front of their own computer can see a virtual conference room with real video captured from the webcam of participants on the monitors. Each position in the virtual conference room can be occupied by the dynamic image of a participant, who indicates his expected seat in the virtual space. By giving gestures as commands, participants can navigate around the conference room and freely change their viewing angle to interact with others. Through the real 3D interactive experience, the proposed system provides a more efficient and interesting communication platform. System design and detailed technique is revealed in Sec. II. Experimental results will be shown in Sec. III. Conclusion is discussed in Sec. IV.

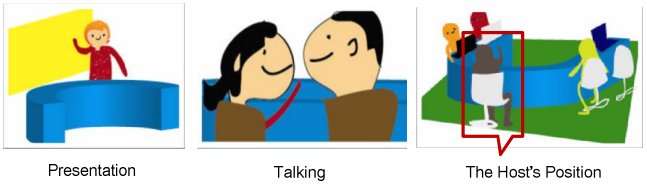


Fig. 1. In the real world, one participant may give a presentation, talk with others, or sit on the host's seat with different mapped positions and orientations in the virtual space.

II. PROPOSED TECHNIQUES

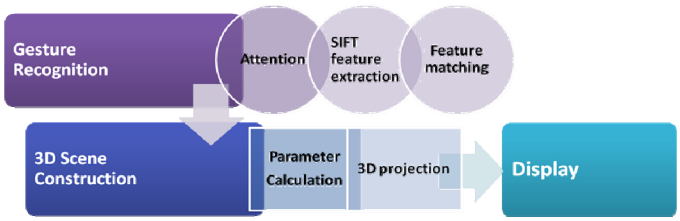
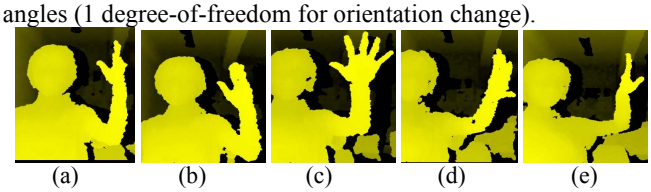


Fig. 2. System flow

Our system provides mix-reality experiences to users, including the scene of the virtual conference room embedded with real dynamic images of participants from remote places. The user can navigate in the virtual space by issuing gesture commands as the human-computer interface. Fig. 2. depicts the approach overview of the proposed system. The system flow can be divided into two parts: gesture recognition and 3D scene construction. The detailed methods are described in the following subsections.

A.Gesture Recognition

In the whole recognition process, we use a 3D sensor to capture depth images of hand gestures. As shown in Fig. 3., depth information provides better performance for hand segmentation even in a cluttered background. Currently, five gesture commands are provided in the proposed system as shown in Fig. 3(a)~(e), representing "turn right with 85-degree", "turn right with 40-degree", "façade", "turn left with 40-degree", and "turn left with 85-degree", to change the viewing angle of the user in the virtual space. In addition, by stretching or retracting arms, the depth change of hands indicates the corresponding advance or backlash movements in the virtual space. In this way, the proposed system provides a 4 degree-of-freedom (DoF) navigation in the virtual conference room, which means that users can move around (3 degree-of-freedom for position change) and freely alter their viewing angles (1 degree-of-freedom for orientation change).



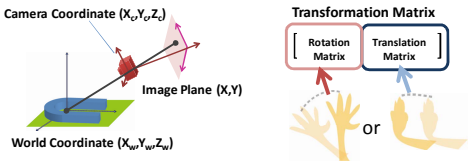
(a) (b) (c) (d) (e)

Fig. 3. Depth images of hand gesture

To recognize the gesture commands, we use SIFT descriptor to recognize different gestures with scale, rotation and luminance invariance. Firstly, the attention region is narrowed down to the area of hand from depth information. Secondly, features are extracted from the interested region. Feature matching is the next step to generate matching pairs by comparing these detected features with features extracted from reference gesture images in the database. Finally, to remove background noise resulting in false matching, the matching pairs are further processed by RANdom SAmple Consensus (RANSAC) algorithm. RANSAC randomly samples a subset of matching pairs and calculate the homography matrix between the reference image and the test image. Then, the matrix is used for all matching pairs to iteratively filter out outliers. The system detects gesture commands of the user at the speed of 15 frames per second and refreshes the scene with different viewing angles according to the current orientation and position of the user.

B. 3D Scene Construction

To project 3D scene to the 2D display of the user according to his current position and orientation in the virtual space, we refer to the pinhole camera model. A camera matrix is used to denote a projective mapping from world coordinates to image plane. Each time the user gives a command to the system, the camera matrix is updated to the current status. Fig. 4(a) shows the spatial relationship among the world coordinate, camera coordinate, and image plane. The camera matrix consists of intrinsic matrix and extrinsic matrix. The intrinsic matrix has parameters related to the focal length, the image format, and the principal point of the used camera. While the extrinsic matrix, which is also called transformation matrix, denotes the coordinate transformation from the 3D world coordinate to the 3D camera coordinate. Based on the camera matrix, the virtual scene is synthesized to reveal real watching experiences for users.



(a) **(b)**

Fig. 4. (a) The camera pose indicates the position and orientation of the user in the virtual space. (b)The transform matrix is composed of rotation and translation matrix, which contains parameters of the viewing angle changes and moving distance, respectively.

III. EXPERIMENTAL RESULTS

We use the PrimeSense, which is adopted on Miscrosoft Kinect, as the 3D sensor to capture the depth image for every gesture command. To measure the recognition accuracy in different gesture commands, we test 40 images and calculate the recognition precision. Table. I. shows that about 90% of recognition rate in average for five categories is achieved. Besides, the detection rate to distinguish between advance and backlash movement is near 95%. The experimental results demonstrate the robustness of the proposed system. Fig. 5. shows that the viewing angle of the user turns left with 40 degrees according to his gesture command.

TABLE I. RECOGNITION RATE OF GESTURE COMMANDS

Command	**Facade**	**R- 40**	**R- 85^0**	**L- 40^0**	**L- 85^0**
Accuracy	95	90.0	87.5	90.0	90

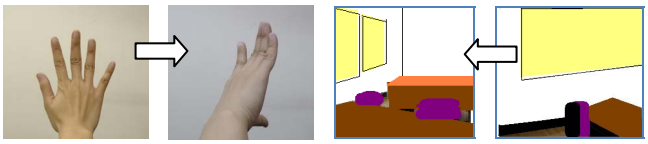


Fig. 5. Gesture and the corresponding viewing angle change

IV. CONCLUSIONS

In this paper, we propose an intelligent 3D online virtual conferencing system with a natural interactive mechanism. The proposed system provides a real 3D interactive experience, which makes communication more efficient and interesting. With the design of the proposed gesture commands, this system supports a nature, simple, and easy way to navigate anywhere in the virtual space and interact with others online. In the future, the proposed gesture commands can be extended for 6DoF navigation, which help people to capture the scene in the virtual space with any viewing angle. Our work can also be applied to many remote applications, like online cooperation, education, concert, team practice, and entertainment.

REFERENCES

[1] WebEX, Cisco corporation, http://www.webex.com/

[2] Peter Kauff et. al, " An immersive 3D video-conferencing system using shared virtual team user environmen", International Conference on Collaborative virtual environments (CVE), 2002, pp. 105-112

[3] Holger Regenbrecht et. al, "An Augmented Virtuality Approach to 3D Videoconferencing", International Symposium on Mixed and Augmented Reality (ISMAR), 2003, pp. 290