MAGIC MUSIC DESK: A MULTI-MODAL EMBODIED INTERACTIVE DESK

Zhiying Zhou, Farzam Farbiz, Xiangdong Chen, Adrian David Cheok, Wei Liu
National University of Singapore Singapore119260

Abstract: *This paper describes a novel multi-modal multi-user audio-visual interface – the Magic Music Desk (MMD) which employs the principles of embodied interaction, and emphasizes social interaction between users. We have developed a novel combination of multiple modalities for the interfaces using speech recognition, hand gesture recognition, sound and visual mixed reality technologies. A new mode of interaction which is called as "What You Say is What You See" (WYSWYS) is demonstrated in our system. This interaction enables all users to visualize each other's spoken words as 3D objects which could be seen by multi-users (this also allows multi-cultural social interaction).*

Key words: Embodied Interaction, Speech, Gesture Recognition, Mixed Reality, 3D-Sound, Multi-Modal Interface

1. Introduction

With the development of new devices for human-computer communication, it has become easier for researchers to design multi-modal (speech, visual, sound, gesture) user interfaces. There has been a large body of research carried on multi-modal interfaces [1] and some prototypes have been already developed [2]. As defined by Coutaz [3], modality refers to the type of communication channel used to convey or acquire information. It also covers the way an idea is expressed or perceived, or the manner an action is performed. In this sense, perceptual user interfaces (PUI) [4], graspable [5] and tangible interfaces [6] can be regarded as systems which

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35660-0_65
R. Nakatsu et al. (eds.), *Entertainment Computing*

have different modalities. One of the important points is that the user interface is moved out of the screen and into the physical environment of the user, or as Ishii described: "to change the world itself into an interface"[6].

Research works that can be regarded as multi-modal interfaces have been reported in mixed reality (MR) category. The Enhanced Desk [7] provides the smooth integration of paper and digital information on a desk and direct manipulation of the digital information with the hands and fingers of users. Rekimoto [8] reported a computer-augmented environment that allows users to smoothly interchange digital information among their portable computers, table and wall displays, and other physical objects. Kato [9] reported an accurate vision-based tracking method for table-top AR environments and tangible user interface (TUI) techniques based on this method. However, few of these research works have examined multi-modal speech, sound, and music interaction in the mixed reality environment.

Embodied computing [10] is a next generation computing paradigm that involves the elements of ubiquitous computing, tangible interfaces and interaction and social computing. In this way, it moves the computer interface away from the traditional keyboard and mouse and into the environment, supporting the more interactive behavior. The three important research paradigms on which embodied computing is founded, are Weiser's ubiquitous computing [11], Ishii's tangible bits or "things that think" [6], and Suchman's sociological reasoning to problems of interaction [12].

Our aim in this research is to develop new interaction modalities which will incorporate these paradigms. We present research that provides ubiquitous computing in a physical environment, tangible interaction (including tangible mixed reality), and emphasizes real-time social computing. Concretely, in this paper we further the field by providing multi-modal interaction system with a new "What You Say is What You See" (WYSWYS) modality in addition to novel speech, music, visual, and sound interactions with an emphasis on social interaction amongst users. The WYSWYS modality is used to visualize speech as a 3D word flowing down to the desk from the speaker's mouth. This allows a new type of social interaction between multi-users, even those who speak different languages. Thus, a multi-cultural social interaction, where words are understood as 3D visual objects can be realised. Additionally, instead of using a "magic paddle" [9] to interact with the augmented virtual objects, we use *both* speech and the real hands to interact with the objects. By these two modalities, users can move the virtual object around the desk or pick up and rotate it as if it is put on your hand. Furthermore, in order to make the MR system more immersive, and to enhance feelings of presence, 3D sound and music is applied according to the movement of augmented virtual objects. We will show that not only is the human-computer interaction multi-modal,

but the interaction between humans also is multi-modal and multi-cultural with this system.

The structure of the remaining part of the paper is as follows: the following section describes the implementation of the Magic Music Desk system and related technologies. In section 3, we describe our real-time application of the Magic Music Desk system, and provide a conclusion in section 4.

2. Implementation of Magic Music Desk and Related Technologies

Figure 1 shows the system architecture of the Magic Music Desk. The top camera is used to capture images for palm and gesture recognition and to give the position of palm relative to the marker. The user's camera is used to follow the user's head movement and to capture the markers on the desk. A microphone is used for recording speech for speech recognition block and earphones playback the 3D sound. The head mounted display (HMD) is used to display the mixed scene which is obtained by augmenting virtual objects on to the real scene.

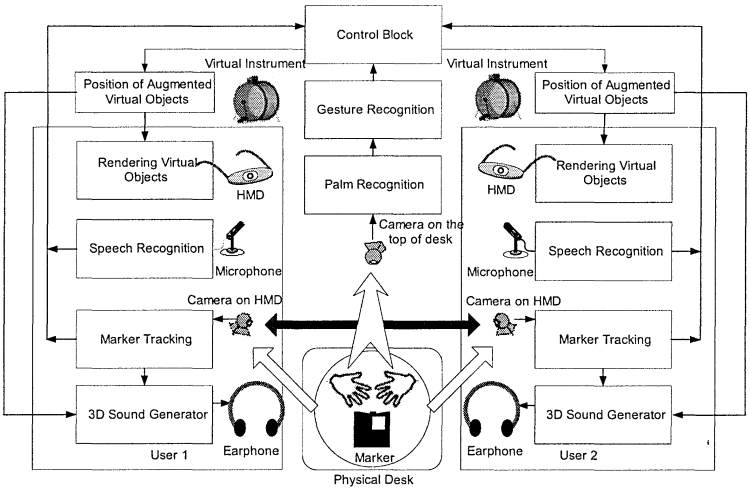

Figure 1. System architecture of the Magic music desk

In our system, speech commands are used to import 3D virtual objects on to the desk and to control objects to move and perform some specific operation such as zooming. The recognized results are sent to the control

block to decide the exact position where the objects should be displayed relative to the markers.

New objects can be imported by simply saying the name of the object like "drum". For more complex commands, we constructed a grammar rule and a model database. From each sentence or command, the speech recognition block will try to extract the object and action. For example, the command "drum rotate left" is separated to "drum", "rotate" and "left", then the recognized results are sent to control block to render appropriate action of the corresponding object.

Stable detection of the palms can be achieved by extracting two kinds of features: statistical-based feature and contour-based feature. Our system recognizes hand gestures with just one camera, and thus avoids the problem of matching image features between different views. Our approach is based on the methods described in [13] and [14] for gesture recognition.

The input images for palm and gesture recognition are captured by the camera on the top of the desk. After the system recognizes the palm, the relative position of palm to the desk (with markers) is able to be calculated (as will be discussed below). This information is sent to the control block to ensure the precise overlay of virtual object on to the palm.

Two simple gestures (Figure 2) are used in our system for the tangible interaction with the virtual objects. When gesture in the top-left figure is recognized and its position is near enough to the virtual object's position, a pick up event occurs. When gesture in the top-right figure is recognized, the object will be "dropped". During the period between picking up and dropping, the virtual object will move with the hand (as long as it is recognized) as if the user is handling the object around. Note that we assume that the hand is near the desk vector plane in all cases of moving the virtual object.

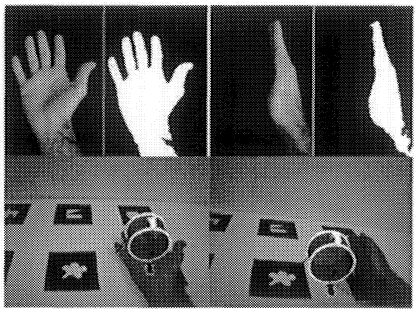

Figure 2. Two gestures used to pick up and drop the virtual object

We employ the marker tracking method [15] to retrieve 3D position and orientation of the marker relative to the camera (given by the transformation

matrix between the marker coordinate and camera coordinate). Figure 3 shows the coordinate systems of the users' cameras and the marker, where T_{C1M} and T_{C2M} denote the transformation matrix between the marker coordinate and camera 1 (user 1) and camera 2 (user 2) coordinates respectively. By the same method we can know the relationship between the marker and the camera on top of desk. Thus, we can combine the 3 cameras in the same coordinate system (marker coordinate) which ensures a precise registration of 3D virtual object.

Thus with these calculations we obtain the three cameras' position in terms of the marker coordinate (i.e. the desk coordinate). As explained above, we obtain the results of palm recognition, moreover, since the camera is fixed on the top of the desk and the markers is also fixed relative to the desk, now we can know the hand's position relative to the markers. This enables our system to know where the palm is and hence to augment the virtual object onto the palm precisely.

Now instead of using a "magic paddle" [9] with an attached tracking symbol to interact with the virtual object, we can use the real hand. As shown in Figure 2, the virtual drum is displayed properly on the user's palm (bottom-left figure) and on the desk (bottom-right figure) while the user drops it.

The marker tracking explained above allow us to calculate the transformation matrix between the camera coordinate and marker coordinate. Thus by individual calculation, we can obtain the transformation matrixes of the two users' cameras relative to the same marker put on the desk.

By shifting the coordinate system from the users' camera and user's mouth (this shifting distance is fixed since the camera is mounted on the HMD), we render a VRML object in a suitable 3D position and thus provide the users an illusion that the words and objects are coming from the mouth. This allows us to produce the novel visualization function of speech, "What You Say is What You See".

In the final function, we provide a novel demonstration of augmented reality by the auditory modality. We employ visual virtual objects to control the simulated direction of the three dimensional (3D) sound. This results in an increased realism of the augmented objects. Here we use sound as another form of tangible interaction with digital objects. The virtual 3D sound becomes another example of tangible interaction.

3. System Results

Here we observe the implementation our system in the form of a Magic Music Desk. Our system enables the user to import and control the virtual

instruments or players simply by using speech commands. In this Magic Music Desk, some instruments and band players are available for user to import and each of them has a related sound. When different combination of the objects is imported on the desk, different music will be heard as if you can "compose" the music by arranging the instruments and players. Note that the sound emanating from each virtual player is fully specialized in the 3D environment. Then the user can use a speech command to move the objects on the table or even use her hand to pick up the object and drop it at the suitable position as she likes. Figure 4 shows that two users are enjoying the Magic Music Desk.

The speech command can be visualized by showing virtual words flowing from the user's mouth. This allows a new mode of human to computer and human to human interaction (WYSWYS). Furthermore even for multi-cultural interaction, the human to human communication can be natural. In Figure 5 we see the example of a user speaking in Mandarin. The word is converted to a 3D virtual character which floats from the user's mouth. Then the object will float down to the table, and "splash down" causing the object that was uttered to appear. Humans can experience an enjoyable cross-cultural and highly visual interaction.

When the instrument is introduced onto the desk, it generates the 3D sound as if the sound is coming from the instrument at that point. Even when you "pick up" the instrument and "move" it, you can feel that the sound source is also moving with your hand. During the whole procedure, speech commands are still available for the user to interact with the objects such as to move it, rotate it, or zoom it.

4. Conclusion

In this paper, we propose a novel multi-modal multi-user audio-visual interface – the Magic Music Desk (MMD) which employs the principles of embodied interaction, and emphasizes social interaction between users. Unlike previous mixed reality interfaces, we implement not only visual mixed reality but also audio mixed reality with more modalities such as speech and 3D sound in a single system. Table 1 summarises the modalities we used in our system.

By speech recognition, our system provides a new method to interact with the augmented virtual objects and provides a novel idea to visualize speech. A novel interaction which is called as "What You Say is What You See" (WYSWYS) is demonstrated in our system. Our system represents a good combination of multiple modalities and enables the user to have a fully-immersive mixed reality environment.

Table 1. Functionalities and interaction with modalities used in Magic Music Desk

Functionalities and interaction	**Modalities used**
Speech Recognition	Speech, audio
Palm and Gesture Recognition	Visual, touch
WYSWYS	Speech, Visual
3D sound	Audio
Tangible Interaction	Visual, Audio, touch

References

[1] C. Mignot, C. Valot and N. Carbonell, "An experimental study of future 'natural' multimodal Human-Computer Interaction", in *Proc. of ACM INTERCHI'93*, Amsterdam, 1993, pp. 67-68.

[2] M. W. Salisbury, "Talk and draw: Bundling speech and graphics", *IEEE Computer*, vol. 23, No. 8, pp. 59-65, Aug. 1990.

[3] J. Coutaz, "Multimedia and multimodal user interfaces: A taxonomy for software engineering research issues", in *Proc. of East-West HCI'92*, St Petersburg, Aug. 1992, pp. 229-240.

[4] M. Turk, "Moving from GUIs to PUIs", in *Proc. of Fourth Symposium on Intelligent Information Media*, Tokyo, Japan, December 1998.

[5] G. W. Fitzmaurice, H. Ishii, and W. Buxton, "Bricks: Laying the foundations for graspable user interfaces", in *Proc. of CHI'95, ACM*, Denver, Colorado, USA, 1995, pp. 442-449.

[6] H. Ishii, B. Andullmer, "Tangible bits: Towards seamless interface between people, bits and atoms", in *Proc. of CHI'97, ACM*, Atlanta, Georgia, USA, 1997, pp. 234-241

[7] H. Koike, Y. Sato, Y. Kobayashi, "Integrating paper and digital information on EnhancedDesk", *ACM Transactions on Computer-Human Interaction (TOCHI)* vol. 8, Issue 4, pp. 307-322, Dec. 2001.

[8] J. Rekimoto and M. Saitoh, "Augmented surfaces: A spatially continuous work space for Hybrid Computing Environments", in *Proc. of CHI'99*, 1999, pp. 378-385.

[9] H.Kato, M. Billinghurst, I. Poupyrev, K. Imamoto, K. Tachibana, "Virtual object manipulation on a Table-Top AR environment", in *Proc. of ISAR 2000*, Oct. 2000, pp. 111-119.

[10] P. Dourish, "*Where the action is: The foundations of Embodied Interaction*", MIT Press, 2001.

[11] M. Weiser, "The computer for the twenty-first century", *Scientific American*, Vol. 265, No.3, pp. 94-104, 1991.

[12] L. Suchman, *"Plans and situated actions: The problem of human-machine communication"*, Cambridge University Press, Cambridge, 1987.

[13] W. Du, H. Li, "Vision based gesture recognition system with single camera", in *IEEE Signal Processing Proceedings*, Beijing, China, vol.2, 2000, pp. 1351 -1357.

[14] L. Gupta, S. Ma, "Gesture-based interaction and communication: automated classification of hand gesture contours". *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 31, No. 1, pp. 114-120, 2001.

[15] H. Kato and M. Billinghurst, "Marker tracking and HMD calibration for a video based augmented reality conferencing system", in *Proc. of IWAR*, San Francisco, CA (USA), 1999, pp85-94.

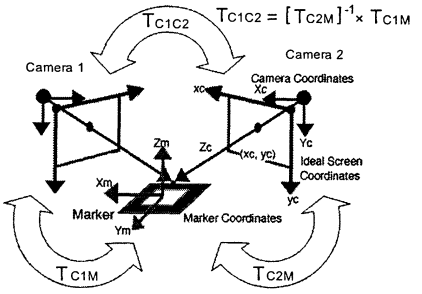

Figure 3. Coordinate systems of users' cameras and marker

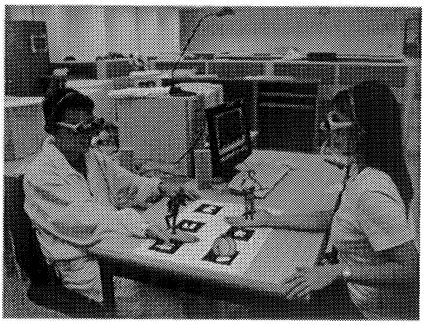

Figure 4. Two users are picking up the virtual musicians and together with the drums, it forms a small band

Figure 5. Visualizing your speech (WYSWYS)