2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
Banff Center, Banff, Canada, October 5-8, 2017

Multimodal Interaction in Augmented Reality

Zhaorui Chen, Jinzhu Li, Yifan Hua
Department of Computing Science
University of Alberta
Edmonton, Canada, T6G 2S4
Email: {zhaorui, jinzhu, yhua3}@ualberta.ca

Rui Shen
VIPShop US
San Jose, CA, USA 95131
Email: rui.shen@vipshop.com

Anup Basu
Department of Computing Science
University of Alberta
Edmonton, Canada, T6G 2S4
Email: basu@ualberta.ca

***Abstract*—With the boost of computing power in mobile devices and availability of cloud APIs in recent years, mobile augmented reality (AR) applications have become increasingly embedded in people's everyday life. However, effective and intuitive interaction between virtual and real worlds in these AR applications is still an open question. In this paper, we make one step towards an answer by exploring the possibility of incorporating two input modalities, gesture and speech, for enhancing user experience in AR applications.**

***Index Terms*—Augmented Reality, Human-Computer Interaction.**

I. Introduction

During the past few years, augmented reality (AR) has changed the way that everyone interacts with machines and with each other. As mobile devices are getting more and more involved in people's daily life, AR applications have become popular in entertainment, navigation, education, retail, and many other areas [1]. While mobile AR (MAR) has shown its great potential in entertainment, current applications are still striving to meet users' increasing demand for more intuitive interaction. According to [2], [3], multiple interaction modalities are needed in an AR application to enhance its user experience. Currently, the two most prevalent input modalities for interaction in MAR are touchscreen and inertial measurement unit (IMU).

Pokemon Go[1], one of the most successful MAR games of 2016, lets a user to toss virtual Poke balls to capture Pokemon, which appear in the AR scene, through the touchscreen of his/her smartphone. Quiver[2], an AR educational drawing application targeting at kids, lets a user to rotate the 3D object created from his/her drawing, through the touchscreen or IMU of his/her smartphone. Zookazam[3], an educational MAR application, and Paparazzi[4], an MAR game, allow users to interact with 3D objects through touchscreen.

While touchscreen and IMU are standard components of a smartphone, other standard input modalities, such as camera and microphone, can also be used for interaction. Paparazzi uses camera to capture the user's eye-gaze and lets the virtual character jump to the gaze point on the screen. This feature makes the character more vivid, compared to other MAR applications, such as Pokemon GO and Zookazam.

[1]http://www.pokemongo.com
[2]http://www.quivervision.com
[3]http://www.zookazam.com
[4]http://pixel-punch.com/project.php?project=Paparazzi

In this paper, we map the two input modalities of camera and microphone to gesture- and speech-based interaction respectively, and explore their effectiveness for communication between the real and virtual worlds. Although here we use a standalone depth-sensing camera (Leap Motion[5]) for gesture input to simplify the processing flow, with 3D sensing capability being gradually built into mobile devices, such as Google Tango[6], we expect that 3D camera would become a standard input modality in the near future.

The rest of the paper is organized as follows. An overview of our AR application is provided in Section II. The two interaction modalities, gesture and speech, are discussed in Section III. Conclusion and future work are presented in Section V.

II. System Overview

In order to explore interaction modalities in AR, we built an AR game, in which a user can interact with a virtual dog. The dog and his living area are superimposed on the real world image, once a predefined marker is captured by the camera. The application was built on Unity engine[7]and Vuforia SDK[8].

The overall structure of our application is depicted in Figure 1. We focus on two interaction modalities, gesture and speech, though other modalities (e.g., gaze) are also supported. The gesture interaction modality is enabled by Leap Motion controller, a depth-sensing device specially designed for real-time hand and finger tracking. The tracking has a high accuracy of 0.7 millimeters [4]. Leap Motion provides a set of gesture control interfaces, such as circle, swipe, key tap and, screen tap.

The speech interaction modality is enabled by Google Cloud Speech API[9], which supports accurate and near real-time speech-to-text translation of 81 different languages. The text is then used for interaction in our application.

III. Interaction Modalities

A. Gesture-based Interaction

Gesture is a natural way for communication, and there are mainly six types of gestures: symbolic, deictic, iconic, pantomimic, beat, and cohesive [5]. Here, we focus on symbolic

[5]https://www.leapmotion.com
[6]https://get.google.com/tango
[7]https://unity3d.com/
[8]https://www.vuforia.com/
[9]https://cloud.google.com/speech

978-1-5386-1645-1/17/$31.00 ©2017 IEEE

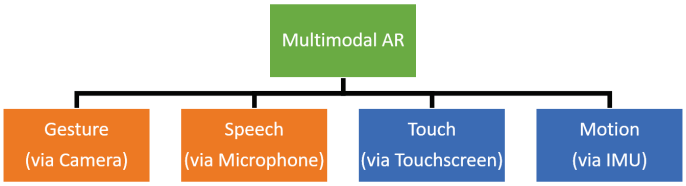

Figure 1: Overall structure of our multimodal AR application. Our application is focused on gesture and speech, which is lacking in most existing MAR applications.

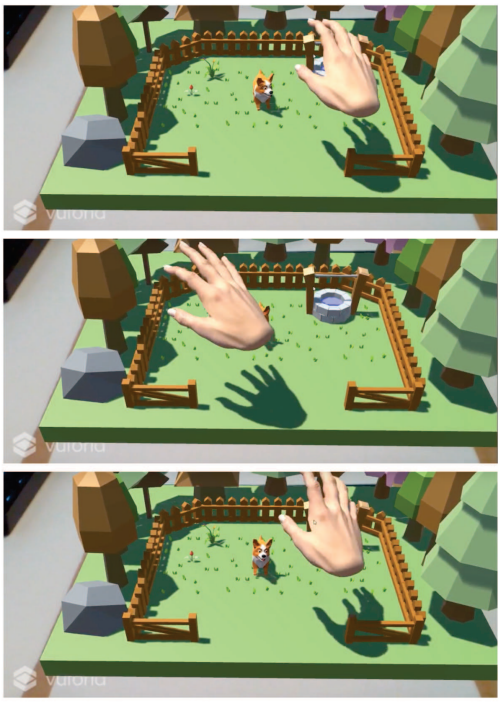

Figure 2: The dog's eyesight and upper body always follow the user's right hand.

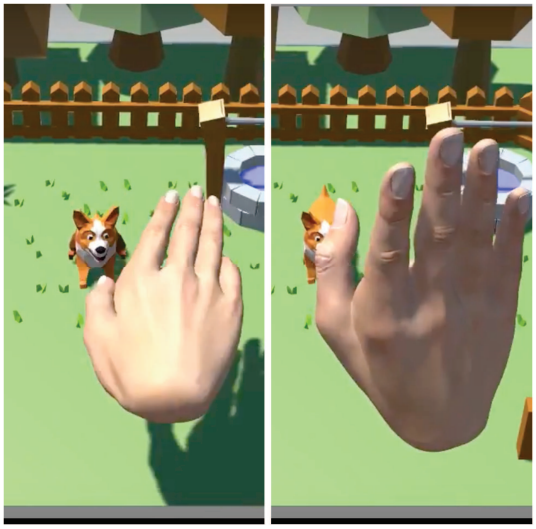

Figure 3: The dog will sit down when the user makes a push gesture (laying down one hand), and stand up when the user makes a pull gesture (raising up one hand).

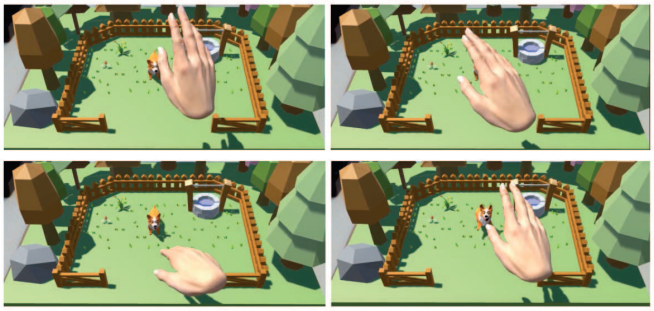

Figure 4: The dog will bark when the user draw a circle, starting from top-left image and ending at bottom-right image.

gestures, each of which conveys a single message. The virtual character in our application is designed to respond to these symbolic gestures from users by performing corresponding actions. This is analogous to how people communicate to or train their pets in real life.

Several gesture-action pairs are currently implemented. A follow gesture is designed as a default gesture. It is called "default" in the sense that the dog's eyesight and upper body continuously follow the user's right hand's movement. Thus, our users are able to wave their hands to interact naturally with the dog, as shown in Figure 2.

In addition, the dog will sit down and stand up when the user makes push and pull gestures, respectively, as shown in Figure 3. The dog will bark when the user draws a circle with one or more fingers and stop barking when user draws a second circle, as shown in Figure 4. A touch gesture, which allows the user to reach out and touch the dog with hands in the virtual world instead of posing gestures in front of it, is also implemented. The dog will lay down in response. This feature provides the user with the opportunity to "feel" the character.

B. Speech-based Interaction

Speech, as a natural interaction modality, has been widely employed in many commercial applications [6]. "The use of gesture is most powerful when combined with other input modalities, especially voice" [5]. Thus, we believe the incorporation of speech modality will bring better user experience in AR. Our application analyzes the user's speech, and let the virtual character perform the corresponding actions.

The interface is shown in Figure 5. On the right-top corner is a button for recording the user's voice input: holding the button for recording and releasing it to stop recording. Below the button is a drop-down menu that includes all 81 languages provided by the Google Speech API. On the left-bottom corner is the log message, showing the result of speech recognition. The four gesture-action pairs shown in Figures 3 and 4 are also supported in this speech interaction modality, as demonstrated

Figure 5: After receiving the user's voice command, the application analyzes the words.

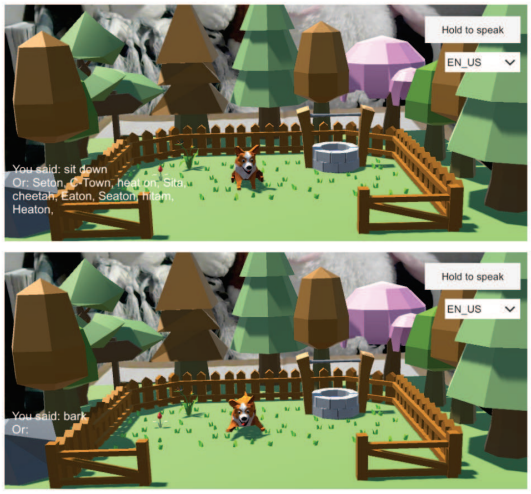

Figure 6: The virtual dog makes response to the voice commands that are defined in the context dictionary. Top: The dog sits down following the sit-down voice command. Bottom: The dog barks when the user says "bark".

in Figures 6 and 7.

In the application, we take advantage of Google Speech API's speech context feature. The speech context is a list of words that could be used as hints for recognition. When developing the application, we input all possible commands that could been said by the users, to the speech context. Then, during the recognition phase, the application first considers the words in the speech context for matching. This greatly improves the correctness of recognition and achieves smoother interaction.

IV. Experimental Analysis

To study the effectiveness of gesture and speech interaction modalities, we performed a preliminary user study.

A. Experimental Setup

Eighteen subjects participated in the experiment. Each subject was asked to perform four tasks (shown in Figure 8) 3 times. In the first round, only gesture commands were allowed. In the second round, only speech commands were allowed. In the third round, a subject was free to use a mixture of gesture and speech commands. A subject was allowed to

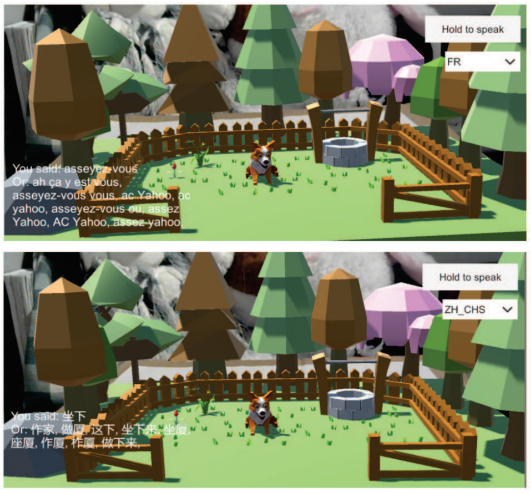

Figure 7: Voice commands in different languages. Top: The dog responds to the French version of sit-down command. Bottm: The dog responds to the Chinese version of sit-down command.

Tasks	Gesture Orders	Speech Orders
Make the dog sit	Push down one hand	Say "Sit down"
Make the dog stand up	Pull up one hand	Say "Stand up"
Make the dog bark	Draw a circle	Say "Bark"
Make the dog stop barking	Re-draw a circle	Say "Stop"

Figure 8: The set of tasks performed by the subjects.

perform multiple trials for each task in each round. A round is considered as completed, when the subject has completed each and every task in the said order. For the round, we recorded elapsed time and computed accuracy. Elapsed time is defined as the duration from starting a command to the completion of the current round. Accuracy is defined as the ratio between the number of successful trials and that of the total trials.

B. Results

The results are summarized in Figure 9, Figure 10, and Figure 11. From these figures, we can see that: 1) speech outperforms gesture in terms of accuracy; 2) speech and gesture have about the same elapsed time; 3) although gesture has relatively lower accuracy, the simplicity of gestures compensates accuracy with speed. Therefore, both gesture and speech are effective interaction modalities for performing simple tasks in AR applications, and by mixing them, one can cover more usage scenarios.

V. Conclusion

In order to explore multimodal interaction in augmented reality, we prototyped an AR application. In addition to common interaction modalities, we added two modalities, gesture and speech, to offer more natural communication between the user and the virtual character. A preliminary user study was performed. As future work, we plan to perform a formal user study to analyze the effectiveness of these two modalities in an mobile AR setup.

Modality	Average Elapsed Time (ms)	Average Successful Task Trials	Average Total Task Trials	Accuracy/Avg
Gesture	37469.44	4	7.72	59.96%
Speech	38727.22	4	4.89	89.35%
Gesture + Speech	38605	4	5.72	79.68%

Figure 9: Comparison of average elapsed time and average accuracy between different interaction modalities.

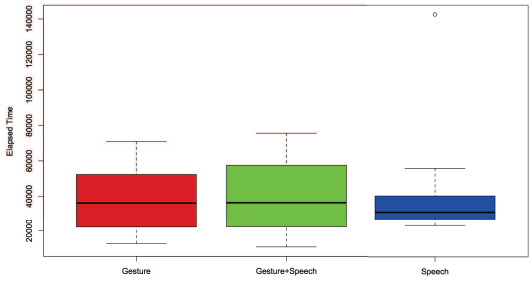

Figure 10: Comparison of elapsed time.

REFERENCES

[1] D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, "Mobile augmented reality survey: From where we are to where we go," *IEEE Access*, 2017.

[2] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer vision and image understanding*, vol. 108, no. 1, pp. 116–134, 2007.

[3] F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, "Human-computer interaction: Overview on state of the art," *International Journal on Smart Sensing and Intelligent Systems*, vol. 1, no. 1, 2008.

[4] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the accuracy and robustness of the leap motion controller," *Sensors*, vol. 13, no. 5, pp. 6380–6393, 2013.

[5] M. Billinghurst, *Haptic Input*, 2011, ch. 14. Gesture based interaction.

[6] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition–a brief history of the technology development," in *Encyclopedia of Language and Linguistics*. Elsevier, 2005.

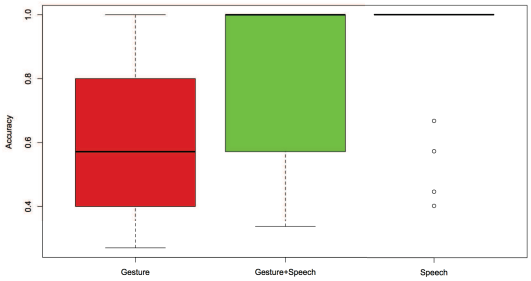

Figure 11: Comparison of accuracy.