Integrating the Kinect Camera, Gesture Recognition and Mobile Devices for Interactive Discussion

Vincent Tam and Ling-Shan Li
Department of Electrical and Electronic Engineering
The University of Hong Kong
Hong Kong SAR
vtam@eee.hku.hk, tholls@hku.hk

***Abstract*—The Microsoft Kinect camera is a revolutionary and useful depth camera giving new user experience of interactive gaming on the Xbox platform through gesture or motion detection. Besides the infrared-based depth camera, an array of built-in microphones for voice command is installed along the horizontal bar of the Kinect camera. As a result, there are increasing interests to apply the Kinect camera for various real-life applications including the control of squirt guns for outdoor swimming pools. In additional to the Kinect camera, mobile devices such as the smartphones readily integrated with motion sensors have been used for different real-time control tasks like the remote control of robots. In this project, we propose to integrate the Microsoft Kinect camera together with the smartphones as intelligent control for interactive discussion or presentation for the future e-learning system. To demonstrate the feasibility of our proposal, a prototype of our proposed gesture recognition and command specification software is built using the C# language on the MS .NET platform, and will be evaluated with a careful plan. Furthermore, there are many interesting directions for further investigation of our proposal.**

***Index Terms*—depth camera; gesture recognition; mobile devices; e-learning systems; interactive discussion**

I. INTRODUCTION

Human-computer interaction (HCI) [6] is one of the main subjects in Information technology and Engineering field. People are constantly searching for a better way to control and to communicate with more convenience and freedom with other people. With the recent advance in gesture-based control technologies and tablet or mobile devices, gesture-based or touch-based input has significantly reshaped interactive discussion or presentation during classes. In a previous work [2], [3], we devised a working mechanism using the precise infrared receiver of a *fixed* Wii Remote [9] together with a BlueTooth-enabled desktop/notebook PC pre-installed with an intelligent software written in C# [10] to localize on the position of a moving infrared pen/emitter, thus allowing *free handwriting* onto the projected images of course notes to facilitate interactive discussion or presentation without any touch-screen.

Nevertheless, the Wii Remote as a flexible sensing device is not hand-free. Nowadays, another modern trend of HCI device is the hand-free movement sensing technology through the infrared-based depth camera. One typical example is the Microsoft Kinect camera that is very popular and successful because of its new gaming experience, the body movement sensing and interaction. The Kinect camera [8] provides the possibility to develop a new kind of HCI methods. In this paper, we propose a radically new working mechanism using the depth camera, intrinsically a pair of precise infrared emitter and receiver, connected to a desktop/notebook PC pre-installed with an intelligent software written in C# to track on the trajectory positions of a moving arm/hand or other body parts, thus allowing *any predefined gesture* to be recognized by a specific machine learning algorithm after some training so as to facilitate interactive discussion or presentation without any touch-screen. More interestingly, the machine-learning algorithm we adapted is both very efficient and effective. It typically requires 3–5 times of trainings for any gesture ranging from the simple to more complicated ones to be recognized with over 90% accuracy. This allows the users or instructors to instantly define any general gesture that can be flexibly mapped into any available command for real-time controls including the lighting, room temperature, other computer applications or even robots.

Especially, this clearly opens up numerous opportunities for various platforms/systems to foster interactive discussion. Our proposed interactive discussion/presentation system is easy to implement, and ready for any course instructor or presenter to use without any need to learn how to write skillfully on the small touch-screens of tablet or mobile devices. Besides the Kinect camera, we have installed a server machine that allows user logins with passwords and personalized gestures to be stored in the back-end database to flexibly define commands for different applications. On top of it, the voice input feature of the Kinect camera can also be used to issue voice command for interactive presentation. Furthermore, our server also supports the use of the very popular quick recognition (QR) codes [12] that can be dynamically generated and sent to the instructor's mobile device via the cellular or WiFi network to turn his/her mobile device into a real-time control device for interactive discussion/presentation. We have demonstrated our working prototype to several students, colleagues and visitors from non-government organizations (NGOs) including the Hong Kong Federation of Youth Groups (HKFYG), who showed very strong interests in adopting our systems for their interactive discussions in the future. A more thorough testing

978-1-4673-2418-2/12/$31.00 ©2012 IEEE

and evaluation of our improved prototype will be conducted around the end of the Fall semester. After all, there are many interesting directions for further investigation including the integration of multimedia files into our improved prototype, the integration of our improved prototype into a simulator or e-learning software, and a thorough study of the pedagogical changes brought by our improved prototype or the integrated system for interactive discussion.

This paper is organized as follows. Section II reviews the system design of our example application in using the Microsoft Kinect™ camera [8], mobile devices such as the smartphone [13] and the QR codes to promote interactive discussion in classes. We give an empirical evaluation of our proposal on various criteria in Section III. Lastly, we summarize our work and shed lights on future directions in Section IV.

II. Our Proposed Interactive Discussion System

Our proposed Kinect-based interactive discussion system consists of 3 main functions, namely the personalized gesture input and recognition subsystem to define user-specific gestures for storage, recognition and retrieval, the voice input and recognition subsystem, and lastly the QR code subsystem. The first two main functions are provided by our back-end cloud server installed with a database to store individuals' gesture and voice patterns mapped to specific commands to applications, aided by a Kinect camera connected to a desktop PC to perform some pre-processing at the front end. The last main function simply involves the direction interaction between our cloud server installed with the QR code generator and the user's mobile device installed with the appropriate QR code reader.

For the gesture input and recognition subsystem, it requires the use of a hidden Markov model as a statistical tool for pattern recognition of the user's gesture inputs as based on the trajectory data returned by the Kinect camera. For the voice input and recognition subsystem, our proposed interactive discussion platform mainly employs a dynamic programming technique, namely the dynamic time warping technique, to match the inputted voice sequence with those stored voice patterns on our cloud server in an optimal way. Lastly, for the QR code subsystem, after successful users' logins, our cloud server will dynamically generate a QR code for sending to the instructor's mobile device installed with the appropriate QR reader to decode the QR code into a webpage link to install a mobile app for turning the mobile device into a real-time control device for interactive presentation.

Fig. 1 shows the basic system architecture of our proposed interactive discussion system in which the Kinect camera is used to capture the real-time trajectory data of the user's hand/arm. The data will be sent to the connected desktop PC to execute a pre-processing application to extract out the essential features from the captured frames by the depth camera. After the extracted features sent through the Internet to our Cloud server, the server will run the designated machine learning algorithm, namely our adapted one-dollar algorithm (as a variant of the hidden Markov model), produce the learning weights for matching against the stored values in the back-end database.

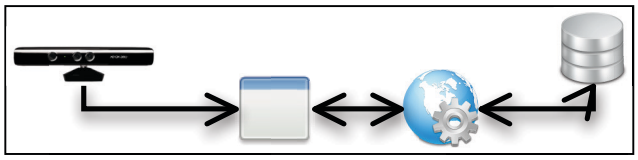

Figure 1. The system architecture of our proposed discussion system.

One of the main uses of our proposed system is to control the PowerPoint slide shows. Some default gestures will be provided for newly registered users so that our system can be used right away. However, most users may prefer to train the cloud server with their own gesture and then assign different actions to various user-defined gestures. After our system is well trained, the user can use those stored gestures to control the PowerPoint slideshows or other applications including other programs or real-time machines like robots. Without the training data set, it can also control the mouse cursor on the screen. If the user prefers, mobile phones can be used as remote controllers through decoding the QR code sent by the cloud server to point to a specific webpage for interactive presentation.

III. Prototype Implementation and Evaluation

To demonstrate the feasibility of our proposal, we implemented a prototype of the Kinect based interactive discussion and presentation system using the C# programming language on the Microsoft .NET platform together with a cloud server machine installed with a library of intelligent machine learning algorithms including the One Dollar recognizer for gesture recognition, or the dynamic time warping (DTW) recognizer [5] for voice recognition, the QR code generator and a back-end database of individually stored gestures. It took around 4 man-months for the design and implementation of our proposed interactive discussion and presentation system.

Fig. 2 shows the user interface of our proposed Kinect-based gesture input and recognition subsystem for course instructors to flexibly define personalized gestures for storage into the back-end database. The default machine learning algorithm is the One Dollar algorithm [14] for gesture recognition due to its best performance obtained from our empirical evaluations as aforementioned. After defining the gestures, course instructors can make use of the interface of the gesture viewer subsystem to map the newly defined gesture into a specific command for controlling certain application such as the forwarding or rollback of the PowerPoint slide shows.

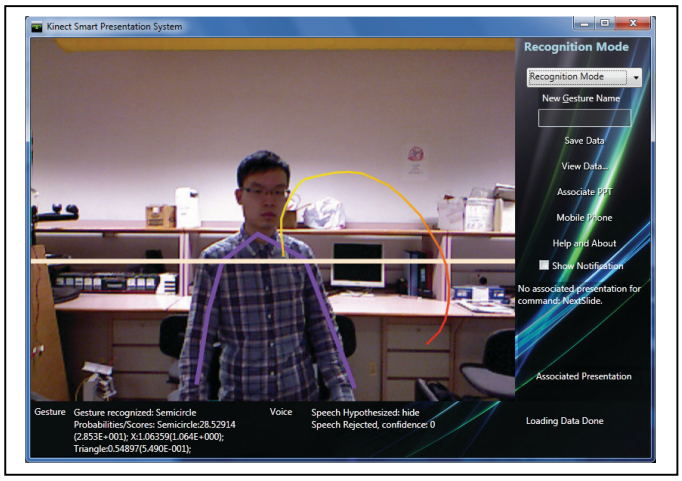

Figure 2. The user interface of our Kinect-based gesture recognition subsystem for defining gestures.

Lastly, our prototype of the proposed interactive discussion system also supports the use of QR code to turn the end users' mobile devices such as the smartphones into real-time control devices for interactive discussion. After the user successfully log into our cloud server, (s)he can request to send a newly generated QR code into his/her smartphone already installed with an appropriate QR reader. When the QR code is decoded on the smartphone or mobile device, it will point to the server's webpage with a carefully designed controller interface to turn the smartphone or mobile device into a real-time control device for interactive discussion / presentation.

IV. Concluding Remarks

In this paper, we devised a radically new working mechanism using the Microsoft Kinect (depth) camera [8] for efficient gesture recognition [11] and flexible command definitions supported by a powerful cloud server installed with intelligent machine learning algorithms like the One Dollar recognizer, or the dynamic time warping technique for voice recognition, as written in the C# language [10] on the Microsoft .NET platform to facilitate interactive discussion or presentation without using any touch-screen. Besides, a database of individuals' gestures is built to allow the storage of personalized gestures after successful users' logins in order to control various real-world applications or hardware including robots. On top of it, the latest QR code related technique and also the controller interface of the web pages are also employed to turn mobile devices like the smartphones into real-time control devices for interactive presentation.

Clearly, our proposal opens up numerous opportunities for various platforms/systems to foster interactive discussion. Our system is easy to implement, and ready for any course instructor or presenter to use without any need to learn how to write skillfully on the small touch-screens of tablet or ultra-mobile PCs. We have demonstrated our working prototype to several students, colleagues or our visitors from the NGOs, who showed strong interests in using it for their interactive discussion in the future. A more thorough testing and evaluation of our improved prototype will be conducted around the end of the Fall semester. After all, there are many interesting directions for further investigation including the integration of multimedia files into our improved prototype, the integration of our improved prototype into a simulator or e-learning software, and a thorough study of the pedagogical changes brought by our improved prototype or the integrated system for interactive discussion.

References

[1] D. J. Baumbach, "Web 2.0 and you," *Knowledge Quest*, vol. 37, no. 4, pp. 12–19, Mar./Apr. 2009.

[2] S. T. Fung, "Using the Wiimote for gesture input and interactive discussion," Dept. Elect. & Electron. Eng., Univ. Hong Kong, Hong Kong, Tech. Rep., 2009.

[3] S. T. Fung and V. Tam, "Towards an innovative application of the Wii remote to facilitate interactive discussion," in *Proc. Int. Conf. ICT in Teaching and Learning*, Hong Kong, 2009.

[4] L.-S. Li, "Smartphones + Kinect = Smart Key Presenter," Dept. Elect. & Electron. Eng., Univ. Hong Kong, Hong Kong, Tech. Rep., 2012.

[5] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.

[6] B. Shneiderman et al., *Designing the User Interface: Strategies for Effective Human–Computer Interaction*, 5th ed.. Boston, MA: Addison Wesley, 2009.

[7] Wikipedia. (2012). *Cloud Computing* [Online]. Available: http://en.wikipedia.org/wiki/Cloud_Computing.

[8] Microsoft. (2012). *Kinect* [Online]. Available: http://www.xbox.com/en-US/kinect.

[9] Wikipedia. (2012). *Wii Remote* [Online]. Available: http://en.wikipedia.org/wiki/Wii_Remote.

[10] Wikipedia. (2012). *C Sharp (Programming Language)* [Online]. Available: http://en.wikipedia.org/wiki/C_Sharp_(programming_language).

[11] Wikipedia. (2012). *Gesture Recognition* [Online]. Available: http://en.wikipedia.org/wiki/Gesture_recognition.

[12] Wikipedia. (2012). *QR Code* [Online]. Available: http://en.wikipedia.org/wiki/QR_code.

[13] Wikipedia. (2012). *Smartphones* [Online]. Available: http://en.wikipedia.org/wiki/Smartphone.

[14] J. O. Wobbrock et al., "Gesture without libraries, toolkits or training: A $1 recognizer for user interface prototypes," in *Proc. ACM Symp. User Interface Software and Technology*, Newport, RI, 2007, pp. 159-168.