OpenGesture: A Low-Cost, Easy to Author, Application Framework for Collaborative, Gesture and Speech Based Learning Applications

Marcelo Worsley
Stanford University

Paulo Blikstein
Stanford University

ABSTRACT

In this paper, we present an application framework for enabling the development of gesture and speech based applications for collaborative learning environments. More specifically, we are concerned with combining the affordances of natural interfaces with educational theories concerning embodied cognition to develop an application framework that enables education researchers and practitioners to create enriching multi-modal learning experiences. Furthermore, this paper highlights a user study that explores how students interacted in a multi-user, collaborative space. Our initial findings indicate that these applications are certainly feasible for well-defined learning tasks, but may require machine learning based training in order to be successful in contexts where the vocabulary is more diverse.

1. INTRODUCTION

The release of the Nintendo Wii and the iPhone ushered in a wave of more natural interfaces, and the proliferation of the iPad and Microsoft Kinect have cemented the salience of these new input modalities as being engaging and exciting for users. While these tools have experienced rapid adoption among consumers, there continues to be a significant lag in the development of educational applications that leverage novel input modalities: gestures, speech, gaze, etc. However, it is in the area of education, that society and individuals may be able to reap the greatest benefits from more natural interfaces, through the affordances of exertion theory and embodied cognition, for example [1][2][3]. That said the slow adoption of those tools in learning environments is not unexpected when one considers the prohibitive expense and the complex authoring environments that they require. Accordingly, we endeavor to develop a multi-modal application framework that is low-cost, feels natural for the user, and can be accessible to a larger community of developers, with a particular focus on making it usable for education researchers and practitioners. More specifically, our system delivers a low-cost, yet high tech, authoring and execution framework for speech and gesture based applications. Furthermore, the framework only utilizes open-source software, and publicly accessible hardware, that is likely to already be available at many homes and schools. Moreover, this framework permits both multi-user and multi-touch interactions, while also enabling researchers and practitioners to capture pertinent multi-modal learning analytics [4][5] about students. To this end the application framework enables the capture of rich user specific information that researchers and practitioners can use to track student interaction with the system. This information includes recording student gestures, speech and video for later analysis.

The following sections briefly describe the system architecture, study methodology and preliminary user study results. These sections are followed with a short discussion of the user feedback. This paper concludes with opportunities for future research and the potential impact of this technology.

2. ARCHITECTURE

At the core of this application framework is the Nintendo Wiimote (a 1024 x 768 infrared camera), an infrared source, a microphone and a web camera. These devices, when combined with a computer, enable low-cost sensing of gestures and audio capture. In addition to these, however, there are a number of other inexpensive technologies that enable increased functionality of the application. The additional items include: additional infrared sources, a large screen display (presumably already available in most education settings) and an array microphone (currently we utilize the Microsoft Kinect for this functionality).

As previously noted, this application framework leverages open-source technology to create a cross-platform tool. The selection of freely accessible software libraries was an intentional design decision that would enable the tool to be adopted by schools and teachers that are already facing shrinking budgets and a range of personal computers and macs. Furthermore, there is nothing about the solution that would preclude a parent or student from setting this up at their residence for a more engaging way of interacting with educational content or media as many families might already own several of the hardware components.

The application framework is built in C/C++, and includes libraries for capturing and manipulating audio, video and Wiimote data, using Sphinx 3, Opencv and PyWii, respectively. For speech recognition, we are using PocketSphinx, a highly extensible open-source tool that permits users to supply their own acoustic models, language models and grammars [6]. Additionally the application framework contains a plugin that makes it easy for users to automatically make language models (a crucial component for making speech based applications). The system

also makes use of LIBSVM [7] for realizing user specific gesture training. Finally, the system is tied together using Qt's Webkit. Webkit allows C++ applications to display HTML content, and also features Javascript integration. Furthermore, Webkit contains DOM integration which allows the application to walk and modify the content of web pages. Thus, by combining the capabilities of Webkit, with speech recognition and gesture recognition software, we are able to make a rich user interface that frees individuals to utilize more natural modalities for interaction. When augmented with techniques from computer vision and speech processing, we are able to analyze collaborative interactions based on user location, and directional speech capture from the array microphone.

3. FRAMEWORK DESIGN

Before discussion the primary extensions involved with this work, we shall briefly describe some of the original design decisions that warrant justification.

Motivation for Using Infrared

With the existing computer vision capabilities exhibited by the Xbox Kinect, one could conceivably question our desire to utilize a seemingly less natural gesturing technique. However, we considered utilizing a hands-free gesturing system for our application framework, but found through initial testing that these hands-free gestures do not necessarily come naturally. Instead, we as a culture have been trained, by way of the television remote, and through the use of writing instruments, to gesture with a prop. In addition to considerations around having a naturalistic gesturing system, the use of an infrared source as opposed to high definition images, was also motivated by a lack of precision in processing hands-free gestures, in addition to the various image segmentation issues required to accurately locate and segment people from images. This second concern about image segmentation is particularly important in dynamic, collaborative spaces, where the additional image noise created by several moving bodies adds additional complications. Using infrared, on the other hand helps eliminate these complications.

That said, because we wish to leave many of these design decisions in the hands of practitioners and researchers, we are currently working to add Xbox Kinect integration for doing gesture capture. By doing so, we permit the application developers to use whichever gesturing technology seems most appropriate to their environment.

User calibration

The initial implementation of our system featured a user calibration step that allowed the user to specify the gesturing region. This process was included to ensure that users would not experience great discomfort or frustration reaching the extremities of the screen. The calibration process also served as a good introduction to the sensitivity and feel of the gesturing system.

User trained gestures

Much like speech patterns, everyone tends to gesture in a slightly different fashion. In order to accommodate this, our system has the capability to take users through a gesture training process for any gestures that they would like to use. This training data is used to make the application more personalized for the individual. Beyond personalization for specific individuals, this capability can allow for more age-consistent gesturing since one can reasonably assume that even something as simple as drawing a circle will vary based on the age of the individuals.

Smart Gesturing – Based on object extent

One of the main benefits of Qt Webkit is the ability to access the DOM for each screen in a given application. By having access to the DOM, we are able to query the extent of each HTML element, which allows for more intelligent gesture recognition. For example, a dwell gesture, need not be tied to a specific x,y pixel location but can instead be tied to a button or an image. In this way, the user need not be concerned about keeping the infrared sensor in the exact same location for the duration of the dwell.

Heuristic for monitoring speech intentionality

The base system contains a heuristic for identifying speaker intent. This heuristic is based on the presence of infrared or a face (as determined by face detection) within 5 seconds of the speech command. Speech commands are also interpreted as being system directed anytime the user is currently activating infrared.

4. PILOT STUDY

This pilot study included approximately 20 high school and college students. These students volunteered to participate in the study. To be clear, the purpose of the study was primarily to examine how students would utilize the speech and gesture capabilities, and to get their feedback on its utility. The inclusion of a specific learning task was merely to ground the study in something concrete that would involve both speech and gesture based input.

The study was conducted with pairs or triads of students in front of a 47 inch LCD display with a webcam mounted to the top of the screen, and the Wiimote affixed to the bottom of the screen. Participants were given an infrared source and a microphone, and provided with brief instructions on the basic capabilities of the system. These basic capabilities included being able to draw on screen by gesturing with the infrared source, navigating through different screens using speech commands and being able to answer questions about geometric progressions, which would require them to speak and gesture. More specifically, students were presented with a series of screens that contained different arrangements of dots. Given those grids, students were asked to come up with all of the possible dot combinations by drawing them on the screen while they counted them. As the task progressed, the students were presented with increasingly different numbers of dots and different grid arrangements, and were ultimately asked how to determine the number of dot combinations without employing the counting technique.

As a part of the introduction to the platform, users were also informed that completing a dwell gesture, of 3 or more seconds, would mimic a mouse click. Most users completed this task in 5-10 minutes, and produced approximately 30-35 multi-modal interactions (gesturing and speaking). After completing the preliminary gesturing task, one user was asked to engage with the other user about how they went about solving the mathematical problems that were posed to them.

As one will observe in the following results sections, the user study produced a wealth of information about the affordances of this type of system for students doing collaborative work.

5. RESULTS

Initial results suggest that while there are still certainly technological features to improve, students consistently identified the merits of such a system for facilitating shared collaborative interactions. More specifically they noted that the capability for each of them to access the entire interface at the same time allowed for more coordinated and fluid group interactions. Moreover, we observed significant increases in student engagement as compared to during more traditional learning activities. This increase in engagement was also articulated by the students as they described the benefits and draw backs of the system.

6. FUTURE WORK

In addition to the mathematics application that we piloted with the 20 students, we recently created two new applications in the area of physics that in large part, leverage existing websites and physics demonstrations. The ease of creating new applications, and the ability to utilize a number of existing websites are two of the primary features of this architecture.

Additionally, we will be conducting additional user testing in the coming weeks, as we have recently upgraded our framework to run more efficiently, and have a larger set of tools to support easy application development.

7. CONCLUSION

In this paper we have described a framework that supports the development of multi-modal application for collaborative learning environments. This framework allows educational practitioners and researchers with minimal programming experience to author rich applications like the Mathematical Imagery Trainer ("M.I.T") [1], for example. It is therefore our hope that this tool will serve as a means for bringing application development of multi-modal learning experiences to a wider audience, while also advancing the area of multi-modal learning analytics.

8. Acknowledgement

Thank you to Claudia Roberts, Michael Yu-Ta and Calvin Fernandez for their work in refactoring the original Python based application into its current C/C++ form. We would also like to acknowledge the contribution of Ibrahim Cotran in implementing some of the gesture recognition algorithms, and to Michael Johnston, for his assistance in developing the first iteration of this tool.

9. REFERENCES

[1] Howison, M., Trninic, D., Reinholz, D., & Abrahamson, D. 2011. The Mathematical Imagery Trainer: from embodied interaction to conceptual learning. In G. Fitzpatrick & C. Gutwin (Eds.), *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*.

[2] Darves, C. & Oviatt, S. 2004. Talking to digital fish: Designing effective conversational interfaces for educational software, in From Brows to Trust: Evaluating Embodied Conversational Agents, Kluwer: Dordrecht, 2004, 271-292.

[3] McKnight, L. and Fitton, D. 2010. Touch-screen technology for children: giving the right instructions and getting the right responses. In: *Proceedings of ACM IDC10 Interaction Design and Children 2010.* pp. 238-241

[4] Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K., (2011). The 2011 Horizon Report. Austin, Texas: The New Media Consortium.

[5] Worsley, M and Blikstein P. (2011). What's an Expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In Proceedings for the 4th Annual Conference on Educational Data Mining.

[6] Huggins-Daines, D, Kumar, M., Chan, A., Black, A.W., Ravishankar, M. and Rudnicky, A.I. 2006. PocketSphinx: A free, real-time continuous speech recognition system for hand-held devices. In: *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing* (ICASSP), May 2006, pp. 185–188.

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/