Towards a distributed framework to analyze multimodal data

Vanessa Echeverría, Centro de Tecnologías de Información - Escuela Superior Politécnica del Litoral, vecheverria@cti.espol.edu.ec
Federico Domínguez, Escuela Superior Politécnica del Litoral, fexadomi@espol.edu.ec
Katherine Chiluiza, Escuela Superior Politécnica del Litoral, kchilui@espol.edu.ec

Abstract: Data synchronization gathered from multiple sensors and its corresponding reliable data analysis has become a difficult challenge for scalable multimodal learning systems. To tackle this particular issue, we developed a distributed framework to decouple the capture task from the analysis task through nodes across a publish/subscription server. Moreover, to validate our distributed framework we build a multimodal learning system to give on-time feedback for presenters. Fifty-four presenters used the system. Positive perceptions about the multimodal learning system were received from presenters. Further functionality of the framework will allow an easy plug and play deployment for mobile devices and gadgets.

Keywords: learning analytics, distributed framework, data synchronization

Introduction

Multimodal learning analytics refers to the development of effective systems to collect, synchronize and analyze data from different communication modalities, to provide on-time feedback. At the bottom of such systems, data gathered from different modalities needs to be reliable in order to analyze learner's behavior and build predictive models that support decision-making through the learning process (Blikstein, 2013).

Nowadays, with emerging technologies and lower costs of devices, a new challenge in multimodal learning analytics has arisen. Data generated by different sensors and devices become harder to manage when trying to capture as much information as possible. Research community has strived to provide fundamentals analysis of data (Scherer, Weibel, Morency, & Oviatt, 2012; Oviatt, 2013; Worsley & Blikstein., 2015); nonetheless, there is a lack of available tools that foster an effortless deployment of such multimodal systems.

In this paper we describe a distributed framework to be used at the top of multimodal systems which helps to: 1) collect and synchronize data through a distributed architecture, 2) manage connections from different devices and sensors; and 3) organize data through recording sessions. This paper is structured as follows: First, we present the related work from former research on gathering and synchronizing data. Then, we explain how the distributed framework architecture was developed. Also, an application example is presented along with an experiment. Finally some discussion about the experience is reported together with further steps related to this work.

Related Work

Research in multimodal data has gained a lot of attention in recent years, independently of the analysis and area to be explored. The central goal of such multimodal systems is to gather data from several sources and analyze data to discover patterns. While most of the work has been done in the analysis of multimodal data from one media source, it is still difficult to find a framework that allows a simpler interconnectivity and ease of data handling and analysis. Manual interactions such as clapping or performing a gesture are common ways that researchers use to start collecting data at the same time (Leong, Chen, Feng, Lee & Mulholland, 2015). Nevertheless, data with imprecise synchronization is the result of this approach. In the literature, we found well-structured software and frameworks to gather data from multiple inputs. These systems allow controlling several inputs through components and translating them into predefined actions or output signals from basic analysis of the data stream (Camurri et al., 2000; Hoste & Signer, 2011). One concern about the mentioned systems is that all input data is processed in the same machine, lacking of scalability to add new inputs.

A framework presented by a research group of the National Institute of Standards and Technology (Diduch, Fillinger, Hamchi, Hoarau, & Stanford, 2008) strives to capture multimodal data from several sources using a decentralized NTP server and one node for each input source. This framework is similar to the one presented in this paper but our approach differs on allowing TCP/IP connections from any input source who wants to subscribe to the capture session.

Framework Architecture

Our framework architecture is based on a publish/subscribe service to synchronize data collection, processing and storage among distributed computational nodes. Data collection is performed by nodes attached to sensors (for example a webcam, microphone, kinect, etc.), depicted as capture device nodes in Figure 1. Each capture device node subscribes to start and stop recording events with the centralized server. These events are triggered by an interface to interact with the system's user. When the event is published, the centralized server starts to synchronize all data coming from device nodes.

At the moment the user triggers a start recording event via the session station, capture device nodes start streaming their raw data to one or more processing nodes (Figure 2). Each processing node handles an input mode, e.g. video, audio, posture, etc., and each capture device node can send one or more streams to several processing nodes (for example the capture device node for the kinect sends several streams to different processing nodes).

All data processing tasks are done in parallel while the session is recording. When the user decides to finish the session, a stop recording event is published and all capture device nodes stop their data streams. Additionally, after this event, the data aggregation service waits for all mode-processing nodes to submit their reports before preparing a feedback summary that is sent to the user (Figure 2).

The purpose of this architecture is to decouple the data processing tasks from the data capture tasks. Capture devices and mode processing nodes can easily be added or removed from the multimodal system without major reconfiguration. Upon registration, each capture device is given one or more Uniform Resource Identifiers URIs of their corresponding processing nodes.

The publish/subscribe server is implemented in a central server using Node.js while all nodes use Python to receive and send events. Messages are not queued or stored and all recorded data is time-stamped locally. All server and node clocks are synchronized using the Network Time Protocol (NTP).

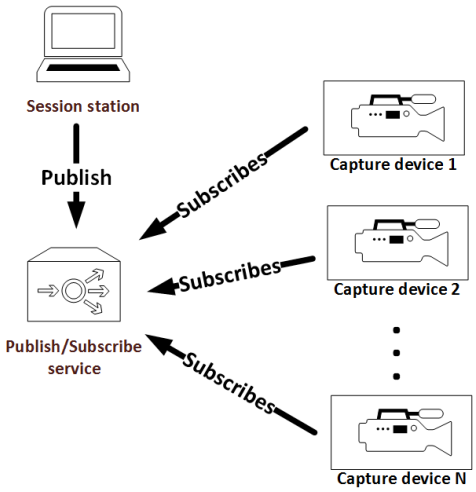

Figure 1: The publish/subscribe service synchronizes multimodal data collection across all capture devices. Each capture device subscribes to the centralized server and starts and stops recording events while a session station publishes these events.

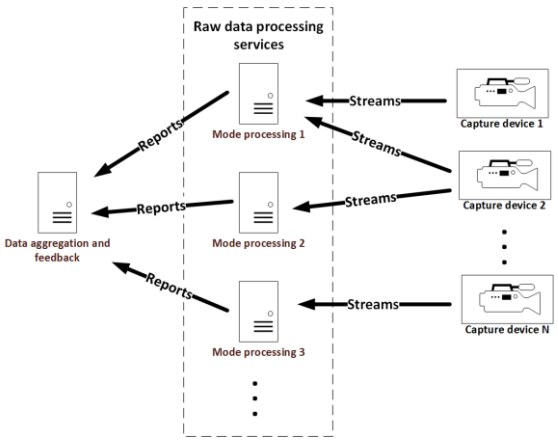

Figure 2: All capture devices stream their data to one or more mode processing services. Each processing service reports its results to a data aggregation service where a feedback is generated and sent to the user.

Application Example: Multimodal Learning System

To test the developed framework, we created a multimodal learning system (MLS) to collect and analyze multimodal data from oral presentation' students. The aim of the MLS is to capture data from several sensors while students present their work orally and to provide on-time feedback at the end of the presentation by analyzing nonverbal skills from gathered data. Thus, we design a physical space to locate all sensors having an immersive, non-intrusive and automatic learning system.

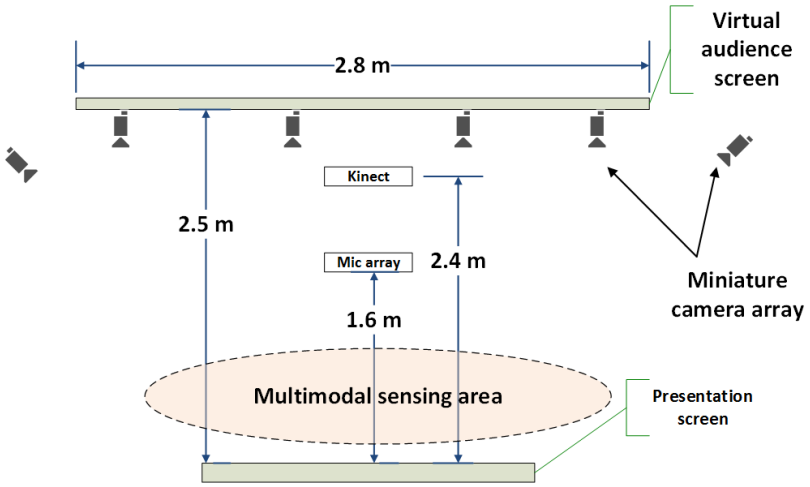

Figure 3. Setup of the multimodal learning system.

Hardware and Software

The MLS is composed of three media streams: audio, video and Kinect data. The audio is recorded using a 6-microphone array with embedded echo cancellation and background noise reduction. This device is located at the lower border of the presenter's field of view. Video is recorded using three Raspberry Pis, each one attached with two low-cost cameras, forming a 6-camera array that covers all sensing area (figure 3 and 4). Kinect data is recorded with a Microsoft Kinect sensor (version 1). This device is located at the lower border of the presenter's field of view, near to the audio device. As depicted in figure 1, all recording hardware is positioned to cover the multimodal sensing area (4 m^2 approximately).

Figure 4. Captured frame from all six cameras.

At the top of the MLS, we created an application using the proposed framework to capture the presenter's data and give on-time feedback to the presenter. The application works as follows: 1) The presenter loads the slides in the computer terminal via USB; 2) The presenter enters a valid email address to receive the feedback results from the application; 3) the presenter starts the recording of all inputs by clicking a "Start" button in the computer terminal; 4) the presenter does the oral presentation meanwhile all the servers gather and analyze each input source; 5) the presenter stops the recording of the data by clicking a "stop" button that appears in the computer terminal; 6) all partial analysis are sent to the central server ; 7) a summary of the analysis is constructed and an email is sent to the presenter.

Data Analysis

Doing oral presentations implies the use of verbal and nonverbal communication skills. The purpose of this MLS is to explore nonverbal skills through the analysis of audio, video and Kinect data streams. Therefore,

from each data stream, a set of features are extracted and analyzed to provide a feedback message after the presentation.

The audio stream is used to measure the clarity of the speech while doing an oral presentation. We calculate the speech rate and detect the filled pauses of the presenter by following the work of De Jong & Wempe (2009) and Audhkhasi, Kandhway, Deshmukh & Verma (2009), respectively.

The video stream from the six-camera array estimates the presenter's gaze. Four of the cameras, located in front of the presenter, indicate if the presenter is looking at the virtual audience screen, while the left and right corner cameras help to point if the presenter is looking at the presentation screen. For each video input the HAAR Cascade face detection algorithm (Lienhart, Kuranov, & Pisarevsky, 2003) is calculated and then, joining all partial results, we obtained the final gaze position, which is determined by one of the two states: facing the audience or watching presentation. At the end, we label each frame with one of the two mentioned states.

Kinect data extracts body posture from skeleton data. Each skeleton frame is composed of 3D coordinates of 20 joints from the full body. For purposes of this application, only upper limbs and torso joints are relevant to calculate body posture of presenter. To determine whether a presenter is doing an specific posture, we define three common postures founded in previous work (Echeverría, Avendaño, Chiluiza, Vásquez & Ochoa, 2014). Thus, the euclidean distances and orientation are calculated from limb to limb at a frame level and, each frame is labeled with one of the three postures.

An additional feature of the presentation was determined by analyzing the presenter's presentation file. Thus, based on a slide tutoring system tool (Echeverria, Guaman & Chiluiza, 2015), we extracted three features from each presentation: contrast, number of words and font size. In the end, the tool determined if the presentation was good or not. The presentation was analyzed per slide and globally.

Feedback

Due to the demanding time when giving on-time feedback in traditional setups, our MLS help us to provide the feedback information right after finishing the presentation. The email that is sent to the presenter shows a summary of the states gathered from each modality. Some predefined messages were inferred after selecting a set of rules that describes whether the nonverbal communication skills were good or bad (figure 5).

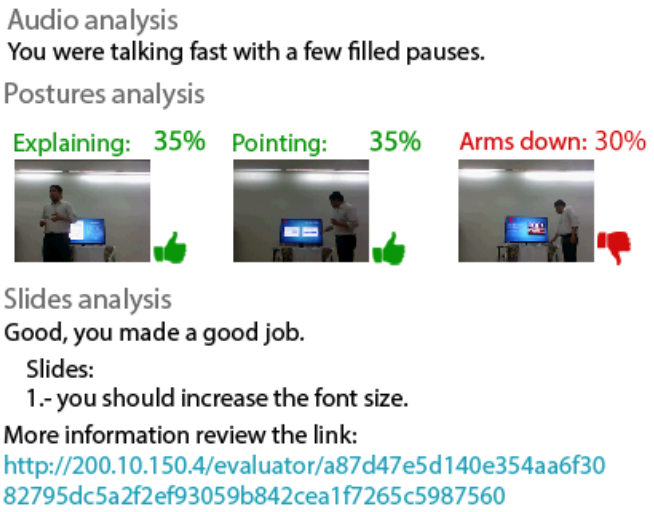

Figure 5. Feedback message received by the presenter. Audio, postures and slides were analyzed.

Thus, for audio analysis we use speech rate and number of filled pauses to determine a set of rules that describe the performance of the presenter while speaking. As for the posture analysis, we took each Kinect frame labeled as: explaining, pointing or arms down and, calculate a percentage for each posture; with this information we also determine a set of rules to describe the presenter's performance according to the body posture. Finally, the results obtained from the slide tutoring system tool helped to create the set of rules for the slide analysis.

Experiment

Fifty-four computer science undergraduate students, 42 male and 12 female, were asked to participate in an experiment to evaluate the proposed framework and the multimodal learning system.

Prior to the presentation, they were informed to select an oral presentation they have previously prepared. The day of the presentation, each student was briefly introduced to the learning system through an explanation on the usage of the system and immediately started the presentation.

Once the oral presentation concluded, the student observed the system's feedback and filled a questionnaire, which consists on six questions using a 10-point likert scale (1: lower value, 10: higher value) and four open-ended questions about the learning system's overall impression and suggestions to improve it.

After recording all students, a manual verification task was carried out to delete presentations where a source was not correctly recorded. Fifty presentations with an average of 8.52 minutes were selected as the final dataset.

Results

Learners' feedback showed positive results about the ease of use, intrusiveness, motivation and experience with non-traditional classrooms (mode: 8), whereas students' perception about the usefulness was reported in a lesser extent (mode: 7). Nonetheless, they think that they learnt anything with the MLS (mode: 9) compared to their previous knowledge. Table 1 shows the minimum, maximum, mode and standard deviation for each likert-scale question.

Table 2: Scores obtained from likert-scale questions.

Question	Min	Max	Mode	Stdv.
On a scale from 1 to 10 with 1 being very awkward, and 10 being very natural, how would you rate your experience with the application?	1	10	8	1.84
On a scale from 1 to 10, with 1 being very motivated, and 10 being very bored, how motivated would you be to use the application again?	1	10	8	2.95
On a scale from 1 to 10, with 1 being low, and 10 being very high, how invasive were the sensors being used to collect data about you?	1	10	1	2.88
On a scale from 1 to 10, with 1 being very likely, and being very unlikely, how likely would you be to use this application in your free time?	1	10	7	2.73
On a scale of 1 to 10, with 1 being not at all, and 10 being completely, do you feel like you learned anything while interacting with the application?	2	10	9	1.98
On a scale of 1 to 10, with 1 being much worse, and 10 being much better, how does using this application compare to how you would normally learn the same content in a traditional classroom?	1	10	8	2.05

From open-ended questions, learners revealed that they learnt from provided feedback specific issues related to posture; slide content and contrast; and filled pauses while speaking.

It is important to note that in the verification task we realized that some of the recordings (sources) were not correctly recorded due to the location of the device according to the presenter's location. For instance, some audio recordings were deleted because of the lower tone of voice; in this particular case, the coverage area of the microphone was overestimated.

Discussion and future work

This paper describes the architecture of a distributed framework to gather and analyze multimodal data. The framework uses a publish/subscribe paradigm to facilitate the connectivity among nodes along with sensors. This framework also helps to maintain all the data well organized and in one place through recording sessions. The analysis of data is made on each dedicated node, which helps to boost the performance of the different algorithms for feature extraction and further analysis.

Using this framework, help researchers to be more efficient to keep all data synchronized. From this experience, we reduced the synchronization time and we put more effort on the analysis of data.

In the future, we will make this framework publicly available. We are going to test this framework not only for mobile devices (e.g. camera/voice recorder from smartphone) but also for digital pens and gadgets or any kind of sensor. In addition, we want to add some functionality such as basic feature extraction algorithms depending on the media to help multimodal community focus on the analysis of data.

References

Audhkhasi, K., Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled-pause detection in spontaneous spoken English. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on (pp. 4857-4860). IEEE.

Blikstein, P. (2013). Multimodal learning analytics. In Proceedings of the third international conference on learning analytics and knowledge (pp. 102-106). ACM.

Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A., Suzuki, K., Trocca, R., & Volpe, G. (2000). Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. Computer Music Journal, 24(1), 57-69.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods, 41(2), 385-390.

Diduch, L., Fillinger, A., Hamchi, I., Hoarau, M., & Stanford, V. (2008). Synchronization of data streams in distributed realtime multimodal signal processing environments using commodity hardware. In ICME (pp. 1145-1148).

Echeverría, V., Avendaño, A., Chiluiza, K., Vásquez, A., & Ochoa, X. (2014). Presentation Skills Estimation Based on Video and Kinect Data Analysis. In Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge (pp. 53-60). ACM.

Echeverria, V., Guaman, B., & Chiluiza, K. (2015). Mirroring Teachers' Assessment of Novice Students' Presentations through an Intelligent Tutor System. In Computer Aided System Engineering (APCASE), 2015 Asia-Pacific Conference on (pp. 264-269). IEEE.

Hoste, L., Dumas, B., & Signer, B. (2011, November). Mudra: a unified multimodal interaction framework. In Proceedings of the 13th international conference on multimodal interfaces (pp. 97-104). ACM.

Leong, C. W., Chen, L., Feng, G., Lee, C. M., & Mulholland, M. (2015). Utilizing Depth Sensors for Analyzing Multimodal Presentations: Hardware, Software and Toolkits. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (pp. 547-556). ACM.

Lienhart, R., Kuranov, A., & Pisarevsky, V. (2003). Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In Pattern Recognition (pp. 297-304). Springer Berlin Heidelberg.

Oviatt, S. (2013). Problem solving, domain expertise and learning: Ground-truth performance results for math data corpus. In Proceedings of the 15th ACM on International conference on multimodal interaction (pp. 569-574). ACM.

Scherer, S., Weibel, N., Morency, L. P., & Oviatt, S. (2012). Multimodal prediction of expertise and leadership in learning groups. In Proceedings of the 1st International Workshop on Multimodal Learning Analytics (p. 1). ACM.

Worsley, M., & Blikstein, P. (2015). Leveraging multimodal learning analytics to differentiate student learning strategies. In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (pp. 360-367). ACM.

Acknowledgments

The authors would like to thank the SENESCYT for its support in the development of this study, to ESPOL's educators and students that participated in the experiment.