Approaches to Multimodal Analysis Of Groups Of Students Learning Online

A. Belucci, L. Cinque, M. De Marsico, S. Levialdi, A. Malizia,
University of Rome "La Sapienza"
Dept of Computer Science
Via Salaria 113, 00198 Roma, Italy
+39-06-8841962
bellucci.andrea@fastwbnet.it,
{cinque, demarsico, levialdi, malizia}
@di.uniroma1.it

S. Tanimoto
University of Washington
Dept of Computer Science and Engineering
Seattle, WA 98195, USA
+1-206-543-4848
tanimoto@cs.washington.edu

ABSTRACT

In this position paper, we describe two existing online learning environments, and we discuss ways in which technologies for audio and video speech analysis might be applied to assessing the learning of online student groups. We present this challenge in the context of mediating and extracting information from multimodal, multiparty online meetings. Online learning communities can have all the attributes of multiparty online meetings, with particular focus on educational agendas.

Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]; K.3.1 [Computer Uses in Education].

General Terms

Multimedia, educational assessment, online learning, meeting analysis.

Keywords

Text analysis, video analysis, educational activity analysis, information fusion, probabilistic reasoning.

1. INTRODUCTION

The promise of online learning is that learning can be more flexible and effective for students in terms of distance, community, and computational tools. For example, a university student might take an advanced course with students from other universities on the subject of computational geometry having affordances for 3D construction and visualization, working in teams of physically separated participants on projects,using state-of-the-art tools for symbolic algebra and proof-finding, and using computer-based laboratory notebooks, as well as video conferencing and online references in the form of web pages and video clips of lectures.

The challenges to online learning include helping students and teachers manage the complexity of the new types of interaction, including forming the teams, synchronizing activities, keeping track of the available resources and finding them when needed, monitoring progress, facilitating communication, fostering creativity, archiving the products of interaction, assessing what has been learned, and relating the group experience to the future needs of the individuals.

First we present two online learning environments. One was developed at the University of Washington in Seattle and is heavily oriented towards archival and assessment of student activity. The other was developed at the University of Rome and is concentrated on multimodal communication in real-time. We discuss the features of these environments for meeting the above challenges, as well as the additional features required for addressing the remaining needs. This results in a possible research agenda for enhancing online learning systems with a new generation of facilities for maximizing the long-term value of collaborative activity in education.

2. THE INFACT LEARNING SYSTEM

At the University of Washington, an online learning environment called INFACT has been developed between 1996 and the present [Tanimoto et al 2000] with the primary purpose of studying automatic techniques for diagnostic educational assessment. INFACT consists of three parts: (1) an online communications forum in which students and teachers can post textual messages with optional graphical sketches, (2) construction tools for students, such as an image processing system and a programming environment for Python, and (3) assessment tools for teachers that allow detection and recording of misconceptions of the student as well as for providing feedback to students.

The most challenging part of INFACT is the third part. Providing automated or partially automated analysis tools to detect student misconceptions typically involves natural

MMMPM'05 Workshop, October 7, 2005, Trento, Italy.

language understanding, sketch understanding, and online activity understanding. These technologies, although more advanced than ever before, remain difficult to implement in new situations. Furthermore, accurate educational assessment requires the fusion of these analysis modalities, and that brings additional design and implementation problems. For an account of some of the challenges, see [Tanimoto et al 2005].

With INFACT, there is an assessment philosophy to keep the assessment process as unobtrusive to the students as possible. This means that multiple-choice testing is normally avoided, and that the evidence for learning is obtained from activity records instead of test answers. Therefore, activities must be designed with both direct learning and assessment in mind, making it more challenging for the designers of the activities. A problem with this methodology is it also begs for a maximum amount of activity data. Currently, INFACT only captures data from students that comes through the keyboard and mouse. No audio or video data are captured, which means that if there are any conversations among co-present students, those conversation streams are lost to the system.

The online learning environment at the University of Rome, described below, on the other hand, mediates video conversations, which means that it has the potential to use video data as part of the activity record of each student.

3. THE ROME COOL-ROOM SYSTEM

At the University of Rome, a collaborative learning environment called CoOL-Room has been developed in order to explore the possibilities for using multimedia technologies and networking in education [De Marsico et al 2004]. CoOL-Room supports synchronous, distance collaborations through a combination of teleconferencing with audio and video and chat and drawing tools. Some notable features of CoOL-Room include a shared drawing space with lockable drawing components, a shared cursor that allows one member at a time to indicate a position on the shared space to the other members of the group, a "radar" view of the shared space that shows who currently controls each drawing component, and the possibility for different roles to be played by different members.

Video communication using small cameras at each group member's location is supported by CoOL-Room. On the other hand, it is also possible for participants to join a session in a "light" mode using text only, either without any camera or with the camera turned off. For typical collaborations, the audio stream is more important than the image stream. This is because unless there is important action or there are manipulable objects involved, most of the information in a session is transmitted through speech. However, in some contexts, such as spoken language learning, video of lip movements can be an important addition to the audio. There are also contexts such as the use of sign language by hearing-impaired people, for which video is essential. CoOL-Room currently provides medium-resolution video communication. However, the video is neither analyzed nor permanently recorded by the system.

4. ANALYSIS METHODOLOGY

4.1 Requirements

In learning environments such as INFACT and Cool-Room, there could be great value in having automatic analyses of audio and video recordings of student communication. In INFACT, these streams could provide missing information needed to perform accurate cognitive diagnoses during online learning. In CoOL-Room, such analysis could add archival value to the representation of a session.

We can identify several needs of these online learning environments. First is to have an intelligible record of each session or activity. This means that basic information must be recorded, such as the start and end times of the session, the names of all the participants, the name or identifier for the chief activity, if known, and whatever details are available about the direct participation of each member. The next need is to have a primary analysis of the session, including some indication of goals being sought in the various stages of the session, an identification of significant events such as introduction of a new goal, communication of a possible solution, discussion of interpretation, gathering and expressing relevant information, and reaching agreements.

Finally, it is desirable if not completely necessary, to have an automatic assessment of what each participant has learned or demonstrated mastery for during the session. These inferences can be difficult to compute, due to limited or uncertain evidence, and so the conclusions should normally be qualified with probability values, assumptions made, etc.

4.2 Information Fusion Using Probabilistic Inference Nets

A possible technology component in the analysis of such streams is the use of probabilistic inference networks. In INFACT, a version of the Bayes Net has been used as a method for combining multiple types of evidence in order to make inferences about student activity and student understanding. This variation, called Sequential Bayes Nets, provides languages for expressing patterns of evidence and for expressing actions to be taken by the system when diagnoses of student misconceptions are made [Tanimoto, Carlson, and Evans, 2005].

When using Sequential Bayes Nets, there are two main phases to the analysis: recognizing evidence and combining evidence. Evidence is recognized in INFACT using string pattern matching on the event description records that are generated as students do their work. For example, if a student posts a message mentioning "curvature" in answer to a question about "shape description", then a rule that matches "shape description" in the subject and "curvature" in the body of the message can fire, activating further processing. The further processing can include combination with other evidence in the fashion typical of graphical models for probabilistic inference. When the probability of a conclusion, given all the evidence, exceeds a threshold, then some suitable action can be taken by the system, such as sending an email message to the student with a suggestion.

5. ANALYZING DIALOG

Current technology in INFACT handles textual message analysis using rule-based pattern matching, in turn using conjunctions of conditions based on regular expressions. INFACT does not yet include provisions for recording or analyzing speech. In order to use speech, two levels of analysis may be tried. At the first, easier level of analysis, speech data would be recorded and attributed to particular speakers, based primarily, if not exclusively, on which microphone recorded the speech. In a distance learning situation, each microphone is associated with a unique user at the time of user log-in. After the speech is recorded, it can be given a superficial analysis that identifies periods of speech activity and inactivity, as well as periods of intense activity. This information can be shown on a timeline, as mentioned later, and it can serve as one kind of evidence in the computation of activity scores and levels of effort.

The second and more ambitious approach to dialog analysis requires the automatic recognition of the speech, and an integration of the semantic representation of the speech with the other information in the system related to the same activities. Even if the meaning of subject-domain terms is not understood by the speech recognition system, the recognition of rhetorically relevant events can be a useful service. The recognition of beginnings and ends of arguments and the separations among the steps of an argument can be useful. In some cases, an educational expert can take the rough analyses from the system and correct or enhance them.

6. ANALYZING VIDEO

Video streams add a new dimension of richness to the session representation. This can enhance the archive both by providing a visual record of the activities and emotions of the students, and by providing additional evidence with which to extract session features such as subgoal setting and achievement. These streams also pose significant challenges for pattern recognition developers.

If there is only one camera per student and the camera is always focused on the student, then many computer vision problems, such as handling occlusions, and person recognition, can be avoided. However, if the online meetings permit more than one person per site, and therefore possibly multiple people in each camera scene, then these problems come back to challenge the vision system. For example, if there are 20 students in an online group, and there are 10 in one location (say a chemistry laboratory) and the other 10 at another location, with one camera per location, then there remains the problem of identifying each person in a scene with 10 at a time, and this is made more difficult when one student walks in front of another, hiding the other, even if only temporarily. Keeping track of the comings and goings of the students within the laboratory, however, could be a useful function for the analysis system.

Even if there is only one camera per participant, there is the possibility that the student moves out of view of the camera, perhaps by getting up and leaving the room during the session. Keeping track of these comings and goings is potentially relevant in this situation, too.

There has been a significant amount of research on the subject of video tracking of moving objects during the past 25 years. One relevant approach uses a combination of background subtraction, segmentation into regions, and tracking the centroids of the regions as they move from frame to frame [Cinque and Malizia, 2005].

Even with this type of tracking performed effectively, there remains an application-related challenge to translate the movements into events that relate directly to such events as activity completion, attention and inattention by the student, and unauthorized obtaining of assistance. Change detection is not the only aspect of such analysis. It is important to validate each user at the beginning of each session, not only by user identification and password, but also by visual presence, and ideally with face recognition. A multi-camera video system for sports event documentation was developed by the Bell Laboratories. Called LucentVision, it was applied to tennis-match capture and broadcast [Pingali et al, 1999]. It illustrates some of the potential for rich video capture. It included facilities for visual analysis of player position and other game-related activities. It did not deal with the analysis of richly semantic audio streams, which would be desirable in a system for educational assessment.

An ultimate challenge for unobtrusive educational assessment is to automatically recognize cognitive state and transitions from one cognitive state to another without direct testing. Video can play a role in the corroboration of hypotheses that come out of the audio analysis. For example, in principle, a student may smile when s/he makes an important discovery in learning. Such “aha” moments are analogous to the times when the crux of a joke is understood. The student finds that some idea or object satisfies two or more pending constraints at the same time. If the video analysis detects a smile, this smiling event thus can influence the probability on an event of greater understanding.

7. DISPLAYING ANALYSES

One of the goals in analyzing meetings in general, and not only educational sessions, is to produce an artifact that helps the participants (and possibly others, as well) to have an overview and a more complete understanding of what happened in the session. The an artifact can serve as the basis for a metacognitive reflection or an inspiration for additional problem-solving activity.

Due to the likely complexity of multimodal data from the meeting, it is important to present the data using clear, easily understood organizational schemes. Such schemes typically give prominence to one or another of the features of meeting items: time of occurrence, responsible member, activity type, or session significance. The INFACT system currently provides a visualization tool to teachers that displays a collection of session events on a timeline. This timeline tool uses the horizontal axis as the time axis and it uses the vertical dimension to spread out events by user and by type.

Important features of timeline tools for meeting analysis and understanding include the following: ability to pan and zoom in time; ability to aggregate events in order to present coherent views at the larger time scales; ability for the user to click on a time point or event icon and get additional details on the event; ability to open up and close various types of events and events belonging to various individuals and groups. INFACT's timeline tool provides implementations for each of these features.

In and when INFACT and CoOL-Room are extended with video capture facilities, there should be timeline facilities implemented to support the video part of the archive. This would include video browsing/playback activated through the timeline, annotation of video clips with temporally placed labels, and the ability to associate students with video clips in such a dynamic way that it lets the teacher see the names of the students in a video clip at any given point in time, for example.

Video-derived events such as smiling, entrances and exits of students from their seats, etc., should have their own stripes on the timeline, with appropriate links to the actual video segments and other evidence that supports them.

8. REMAINING DIFFICULTIES

Even when the algorithms and computational structures for filtering evidence and propagating probabilities are designed and implemented, there remain some difficult problems for designers and users of these online learning systems with enhanced archiving. One problem is that the pattern recognition rules and inference networks have to be updated as the educational content and the student activities are changed. This can be difficult even for experienced teachers. The other problem is that the students may become more and more confused and suspicious as the complexity of the assessment system grows. This points for a need to design and study new techniques for making educational assessment techniques "transparent."

The first problem begs for methodologies for designing evidence-fusion models for particular applications. Sometimes, the inference problems are so difficult that the educational activities themselves need to be modified. Students may need to be trained to be more expressive, so that they provide more verbal and nonverbal clues to the system about what they are thinking.

The other challenge is one of making complex systems more understandable to end users. This can be called the "transparency problem for complex information systems." Transparency can be addressed by a number of techniques including openness of software, user modeling and presentation of qualitative models of the system, and translation of system subcomponent activity into explanations of system behavior.

When it is multiparty meetings that are being assessed, the results of the assessment may be even more complex than for individuals. Not only do the individual contributions and viewpoints need to be determined and represented, but the interactions among the people also need to be determined and recorded at appropriate levels of description. In educational communities, key interactions are those in which some of the individuals contribute ideas or behavior to the group experience that lead others to eliminate a misconception or improve their understanding. Automatically detecting these events will remain an important challenge for the foreseeable future.

9. ACKNOWLEDGMENTS

Our thanks to the sponsors of the research and the organizers of the workshop. Recent sponsorship of INFACT development has been by the U.S. National Science Foundation under grant EIA-0121345.

10. REFERENCES

[1] Cinque, L., and Malizia, A. 2005. A real-time open source video tracking and recognition system. Unpublished draft paper. Dept. of Computer Science, University of Rome "La Sapienza".

[2] De Marsico, M., Fratarcangeli, S., Levialdi, S., and Lombardo, L. 2004. CoOL-Room: Collaboration oriented learning room. *Proc. 2004 Symposium on Visual Languages and Human Centric Computing VLHCC'04*, held at Rome.

[3] Pingali, G., Jean, Y., and Carlbom, I. 1999. LucentVision – A system for enhanced sports viewing. *Proc. Visual99*, (June). pp. 689-695.

[4] Tanimoto, S., Carlson, A., and Evans, N. 2005. Educational assessment using Sequential Bayes Nets. Unpublished paper, Dept. of Computer Science and Engineering, University of Washington. Seattle, WA.

[5] Tanimoto, S., Carlson, A., Hunt, E., Madigan, D., and Minstrell, J. 2000. Computer support for unobtrusive assessment of conceptual knowledge as evidenced by newsgroup postings. *Proc. ED-MEDIA 2000*, Montreal, Canada, June.

[6] Tanimoto, S., Hubbard, S., and Winn, W. 2005. Automatic textual feedback for guided inquiry learning . *Proc. AIED'05*, held at Amsterdam. (July) pp.662-669.