Chapter 11: Multimodal Learning Analytics

Xavier Ochoa

Escuela Superior Politécnica del Litoral, Ecuador

DOI: 10.18608/hla17.011

ABSTRACT

This chapter presents a different way to approach learning analytics (LA) praxis through the capture, fusion, and analysis of complementary sources of learning traces to obtain a more robust and less uncertain understanding of the learning process. The sources or modalities in multimodal learning analytics (MLA) include the traditional log-file data captured by online systems, but also learning artifacts and more natural human signals such as gestures, gaze, speech, or writing. The current state-of-the-art of MLA is discussed and classified according to its modalities and the learning settings where it is usually applied. This chapter concludes with a discussion of emerging issues for practitioners in multimodal techniques.

Keywords: Audio, video, data fusion, multisensor

In its origins, the focus of the field of learning analytics (LA) was the study of the actions that students perform while using some sort of digital tool. These digital tools being learning management systems (LMSs; Arnold & Pistilli, 2012), intelligent tutoring systems (ITSs; Crossley, Roscoe, & McNamara, 2013), massive open online courses (MOOCs; Kizilcec, Piech, & Schneider, 2013), educational video games (Serrano-Laguna & Fernández-Manjón, 2014), or other types of systems that use a computer as an active component in the learning process. On the other hand, comparatively less LA research or practice has been conducted in other learning contexts, such as face-to-face lectures or study groups, where computers are not present or have only an auxiliary, not-defined role. This bias towards computer-based learning contexts is well explained by the basic requirement of any type of LA study or system: the existence of learning traces (Siemens, 2013).

Computer-based learning systems, even if not initially designed with analytics in mind, tend to capture automatically, in fine-grained detail, the interactions with their users. The data describing these interactions is stored in many forms; for example, log-files or word-processor documents that can be later mined to extract the traces to be analyzed. The relative abundance of readily available data and the low technical barriers to process it make computer-based learning systems the ideal place to conduct R&D for LA. On the contrary, in learning contexts where computers are not used, the actions of learners are not automatically captured. Even if some learning artifacts exist, such as student-produced physical documents, they need to be converted before they can be processed. Without traces to analyze, computational models and tools used traditionally in LA are not applicable.

The existence of this bias towards computer-based learning contexts could produce a streetlight effect (Freedman, 2010) in LA. This effect takes its name from a joke in which a man loses his house keys and searches for them under a streetlight even though he lost them in the park. A police officer watching the scene asks why he is searching on the street then, to which the man responds, "because the light is better over here." The streetlight effect means looking for solutions where it is easy to search, not where the real solutions might be. The case can be made for early LA research trying to understand and optimize the learning process by looking only at computer-based contexts but ignoring real-world environments where a large part of the process still happens. Even learners' actions that cannot be logged in computer-based systems are usually ignored. For example, the information about a student looking confused when presented with a problem in an ITS or yawning while watching an online lecture is not considered in traditional LA research. To diminish the streetlight effect, researchers are now focusing on how to collect fine-grained learning traces from real-world learning contexts automatically, making the analysis of a face-to-face lecture as feasible as the analysis of

a MOOC session. More contemporary works on LA explore the new sources of data apart from traditional log-files: student-generated texts (Simsek et al., 2015), eye-tracking information (Vatrapu, Reimann, Bull, & Johnson, 2013) and classroom configuration (Almeda, Scupelli, Baker, Weber, & Fisher, 2014) to name a few. The combination of these different sources of learning traces into a single analysis is the main objective of multimodal learning analytics (MLA).

MLA is a subfield that attempts to incorporate different sources of learning traces into LA research and practice by focusing on understanding and optimizing learning in digital and real-world scenarios where the interactions are not necessarily mediated through a computer or digital device (Blikstein, 2013). In MLA, learning traces are combined from not only extracted from log-files or digital documents but from recorded video and audio, pen strokes, position tracking devices, biosensors, and any other modality that could be useful to understand or measure the learning process. Moreover, in MLA, the traces extracted from different modalities are combined to provide a more comprehensive view of the actions and the internal state of the learner.

The idea of using different modalities to study learning, while new in the context of LA, is common in traditional experimental educational research. Adding a human observer, which is by nature a multimodal sensor, into a real-world learning context is the usual way in which learning in-the-wild has been studied (Gall, Borg, & Gall, 1996). Technologies such as video and audio recording and tagging tools have made this observation less intrusive and more quantifiable (Cobb et al., 2003; Lund, 2007). The main problem with the traditional educational research approach is that the data collection and analysis, due to their manual nature, are very costly and do not scale. The data collection needs to be limited in both size and time and data analysis results are not available fast enough to be useful for the learners being studied. If different modalities could be recorded and learning traces could be automatically extracted from them, LA tools could be used to provide a continuous real-time feedback loop to improve learning as it is happening.

As would be expected, extracting learning traces from raw multimodal recordings is not trivial. Techniques developed in computer vision, speech processing, sketch recognition and other computer science fields must be guided by the current learning theories provided by learning science, educational research, and behavioural science. Given its complexity, the MLA subfield is relatively young and unexplored. However, initial studies and early interdisciplinary co-operation between researchers have produced positive results (Scherer, Worsley, & Morency, 2012; Morency, Oviatt, Scherer, Weibel, & Worsley, 2013; Ochoa, Worsley, Chiluiza, & Luz, 2014, Markaki, Lund, & Sanchez, 2015). This chapter is an initial guide for researchers and practitioners who want to explore this subfield. First, the main modalities used in MLA research will be presented, analyzed, and exemplified. Second, the real-world settings where MLA has been applied are studied and classified according to their main modalities. Finally, several unresolved issues important for MLA research and practice are discussed.

MODALITIES AND MEDIA

In its communication theory definition, multimodality refers to the use of diverse modes of communication (textual, aural, linguistic, spatial, visual, et cetera) to interchange information and meaning between individuals (Kress & Van Leeuwen, 2001). The media – movies, books, web pages, or even air – are the physical or digital substrate where a communication mode can be encoded. Each mode can be expressed through one or several media. For example, speech can be encoded as variations of pressure in the air (in a face-to-face dialog), as variations of magnetic orientation on a tape (in a cassette recording), or as variations of digital numbers (in an MP3 file). As well, the same medium can be used to transmit several modes. For example, a video recording can contain information about body language (posture), emotions (face expression), and tools used (actions).

By its own nature, learning is often multimodal (Jewitt, 2006). A human being can learn by reading a book, listening to a professor, watching a procedure, using physical or digital tools, and any other mode of human communication where relatively complex information can be encoded. The learning process also uses several feedback loops – for example, a student nodding when the instructor asks if the lesson was understood, or the emphasis of the instructor's voice while explaining a topic. These feedback modes usually encode simpler information but are critical for the process. If learning is to be analyzed, understood, and optimized, traces of the interactions occurring in each of the relevant modes should be obtained. MLA focuses on extracting these traces from the different modes of communication while being agnostic of the medium where those modes are encoded or recorded.

The following subsections present the state-of-the-art on the capture and trace-extraction for the most common modalities used in MLA research. For each modality, a brief definition is presented, together with a discussion of its importance to understanding the learning process, a list of most common methods of capture and recording, and examples of where they

have been used. This is not a comprehensive list of all the modes relevant for learning, only those used successfully in MLA studies.

Gaze

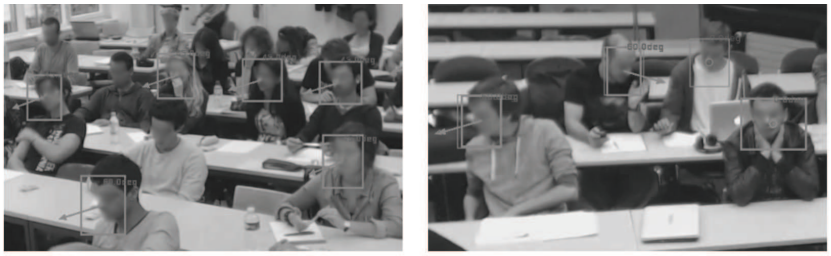
Humans tend to look directly at what draws their attention. As such, the direction of the gaze of an individual is a proxy indicator of the direction of his or her attention (Frischen, Bayliss, & Tipper, 2007). Attention is an indispensable requirement for learning (Kruschke, 2003). Paying attention to a signal helps the individual to capture its information and store the relevant parts in long-term memory. While gaze is not the only proxy to estimate attention and is not error-free, it is commonly used in educational practice. For example, a trained instructor can assess the level of attention of a whole classroom by surveying the gaze of the students; an observer can determine a participant's level of attention in a discussion by tracking the re-direction of the gaze from speaker to speaker.

The importance of gaze has been long identified by marketers, behavioural, and human–computer interaction researchers. Eye-tracking studies are common to determine the effectiveness of advertising (Krugman, Fox, Fletcher, Fischer, & Rojas, 1994), help with the early diagnosis of autism (Boraston & Blakemore, 2007), and the effectiveness of computer interfaces (Poole & Ball, 2006). However, the main methods for recording gaze in these studies, using monitor fixed eye-trackers or special eye-tracking glasses, are too intrusive and costly to be widely deployed in learning settings. The current medium of choice for gaze capturing in MLA is video recordings (Raca & Dillenbourg, 2013). A camera, or an array of cameras, is positioned to record the head and eyes of the subject(s). Then, computer vision techniques, such as those presented in Lin, Lin, Lin, and Lee (2013), are used to extract the gaze direction information from the video recording. The main aspects that need to be controlled to obtain the relative gaze direction in the recording are face resolution and avoiding occlusion from objects or other individuals in the setting (Raca & Dillenbourg, 2013). Information about the position of the cameras in the learning setting must also be recorded to calculate the absolute gaze direction.

MLA has several examples of gaze trace extraction. Raca and Dillenbourg (2013) estimate gaze direction from head orientation in video recordings of students sitting in a lecture using a part-based model (Figure 11.1). In this figure, student faces are automatically recognized (rectangle) and their gaze (arrow) is estimated based on a human face model. This information is then used to determine the focus of attention of individual students and compare it with self-reported attention. Raca and Dillenbourg found that the percentage of time students have the instructor in their field of vision is an important predictor of the level of attention reported. In a different learning setting, Echeverría, Avendaño, Chiluiza, Vásquez, and Ochoa (2014), also estimated gaze direction measuring head orientation by calculating the distance between eye centre points to nose tip point. This information was used to determine if students maintained eye contact with the audience during academic presentations.

Posture, Gestures, and Motion (Body Language)

Posture, gestures, and motion are three interrelated modes, jointly referred as body language, although each one could carry different types of information (Bull, 2013). Posture refers to the position that the body or part of the body adopts at a given moment in time. The posture of a learner could provide information about their internal state. For example, if a student is seated with the head resting on the desk, the instructor could infer that the student is tired or not interested in the lecture. In special cases, the posture adopted is related to the acquisition of skills. For example, students training in oral presentations are expected to use certain postures (hands and arms slightly open) rather than others (hands in the pockets). Gestures being learned do not indicate an internal state. Gestures are coordinated movements from different parts of the body, especially the head, arms, and hands to communicate a specific meaning. This non-verbal form of communication is usually conscious. It is used as a way to provide short feedback loops and alternative emphasizing channels in the learning process. For example, the instructor pointing to a specific part of



a) b)

Figure 11.1. Gaze estimation in a classroom setting (Raca, Tormey, & Dillenbourg, 2014).

the blackboard or a student raising his shoulders when confronted with a difficult question. Finally, motion is any change in body position not necessary to acquire a new posture or to perform a given gesture. This motion is often the result of unconscious body movements that reveal the inner state of the subject during the learning process; for example, erratic movements that signal nervousness or doubt.

Posture, gestures, and motion have been the modes most often studied in MLA due the relative ease in capturing video in real-world environments, together with the availability of low-cost 2-D and 3-D sensors and high-performing computer vision algorithms for feature extraction. While body language can be captured with high precision using accelerometers attached to different body parts (Mitra & Acharya, 2007) or using specialized tools (for example, a Wii Remote; Schlömer, Poppinga, Henze, & Boll, 2008), in practice using them is too invasive or foreign in most learning activities. The most common solution to capture motion is recording video of the subject and estimating posture, gestures, and motion. Any type of camera can be used as long as it can capture the relevant motion with enough resolution. The resolution needed depends on the type of feature extraction conducted with the video. For automatic extraction of human motion, the most common device used is the Microsoft Kinect (Zhang, 2012). Through a mixture of video and depth capture, Kinect is able to provide researchers with a reconstructed skeleton of the subject for each captured frame, which is ideal for capturing body postures and gestures. Newer versions of the Kinect sensor are also able to extract hand gestures (Vasquez, Vargas, & Sucar, 2015).

The most salient examples of the capture and processing of body language in MLA are the estimation of attention through upper-body relative movement delay in a classroom setting (Raca, Tormey, & Dillenbourg, 2014) and the posture and gesture analysis of a novice academic presenter towards the creation of an automated presentation tutor (Echeverría et al., 2014). Figure 11.2 presents the 23 different postures

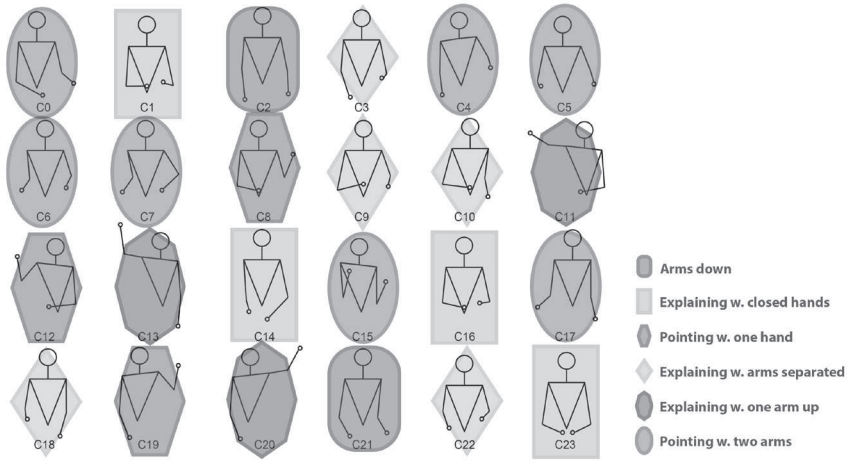


Figure 11.2. Clustered upper-body postures of real student presenters (Echeverría, Avendaño, Chiluiza, Vásquez, & Ochoa, 2014).

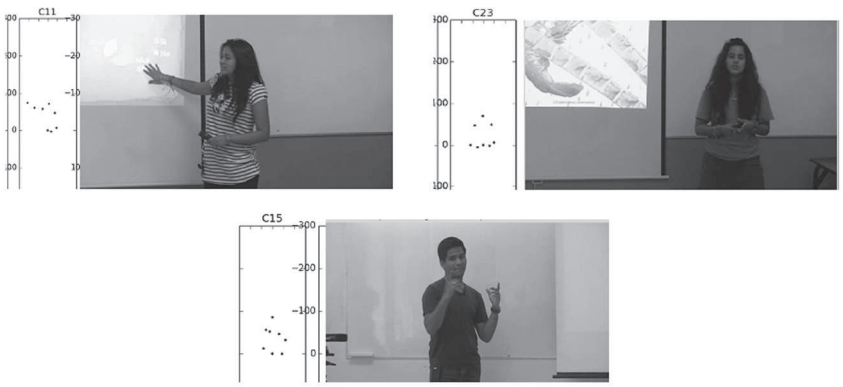


Figure 11.3. Actual postures classified according to prototype postures (Echeverría et al., 2014).

obtained from the analysis of Kinect data of students presenting their work. These 23 postures were classified into six body gestures (different colours) that could be considered good or bad for a presentation. Figure 11.3 presents real examples of these body gestures during actual presentations. The classification of the pose (above the Kinect points on the left) corresponds with what a human observer could interpret from the photo (on the right).

Other interesting examples in using gestures are Boncoddo et al. (2013), Alibali, Nathan, Fujimori, Stein, and Raudenbush, (2011), and Mazur-Palandre, Colletta, and Lund (2014). In the first, Boncoddo et al. (2013) captured the number of relevant gestures performed during the explanation of mathematical proofs and established the relation with the way students arrive at their answers. In the second, Alibali et al. (2011) classified the different gestures made by teachers during math classes and found relations between them. Finally, Mazur-Palandre et al. (2014) presented a study on the use of gestures by children when explaining procedures and instructions.

Actions

The action mode is very similar to the gesture and motion modes. Both are body movements usually captured by video recordings in MLA. However, actions are purposeful movements, usually involving the manipulation of a tool, that are usually learned. The type, sequence, or correctness of these actions can be used as indicators of the level of mastery that the learner has achieved in a given skill. For example, the order and security in which diverse tools are manipulated by a student in a lab can be used as a proxy to determine the understanding that the student has about a given procedure.

The main uses of action recording and analysis in MLA are in expertise estimation. In an engineering building activity, for example, the analysis of hand and wrist movement can determine the level of expertise (Worsley & Blikstein, 2014b). In mathematical problem solving, the percentage of time that a learner uses a calculator can be measured (Ochoa et al., 2013). Ochoa et al. (2013) tracked the position and angle of the calculator in problem-solving sessions (Figure 11.4). This position and angle (line) were then used to estimate which student was using the calculator during that specific frame in the video (intersection with the border of the image).

Facial Expressions

Also highly related to body language modes is the information gathered through facial expressions. The human face can communicate very complex mental states through relatively simple expressions. There has been a large body of successful research in the area of computer vision, trying to identify emotions automatically from facial expressions recorded in video (Mishra et al., 2015).

The main examples of using facial expressions in the field of LA are the works of Craig, D'Mello, Witherspoon, and Graesser (2008), and Worsley and Blikstein (2015b). Craig et al. (2008) automatically estimated the emotional states of students while using the AutoTutor system (Graesser, Chipman, Haynes, & Olney, 2005). Worsley and Blikstein (2015b) used similar techniques to discover emotional changes when students are confronted with different building exercises. Both studies discovered that a confused expression is a good indicator of the success of the learning process.

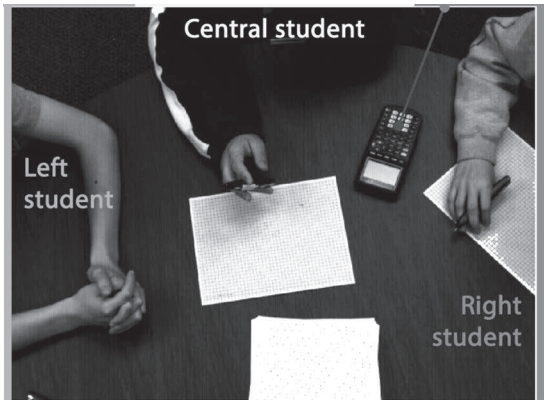


Figure 11.4. Determination of calculator use for expertise estimation (Ochoa et al., 2013).

Speech

The most common use of audio recordings in MLA is to capture traces of what the student is talking about or listening to. As the main and most complex form of communication among humans, speech is especially important in understanding the learning process. In the current practice of MLA, two main signals are extracted from audio recordings: what is being said and how it is being said. In the first approach, usually referred as speech recognition, the actual content of speech is extracted. The result of this analysis is a transcript that can be further processed using natural language processing (NLP) tools to establish what the subject is talking about. In the second approach, prosodic characteristics of the speech, such as intonation, tone, stress, and rhythm, are extracted. These characteristics can shed light on the internal state (security, emotional state, et cetera) or intention of the speaker (joke, sarcasm, et cetera). Speech recognition is heavily dependent on the language used, while prosodic characteristics are less sensible to language variations.

Audio is captured through microphones. While it

seems easier to capture than video, the recording of audio of high enough quality to be processed is actually much more complicated. The type and spatial configuration of the microphones depend on the learning environment and what type of analysis will be conducted with the recorded signal. For example, if automatic speech recognition will be attempted, the microphone should be directional and be close to the subject's mouth. On the other hand, if only the detection of when somebody is talking is needed, an environmental microphone located in the middle of the room could be enough. The presence of noise and multiple signals not only prevents automatic feature extraction but will also degrade manual annotation. The most common technique used to improve recordings, when individual close-recording is not possible, is the use of microphone arrays that can not only reduce the noise but also determine the spatial origin of the audio.

Due to its importance, audio is also present in most MLA works to date. Different types of speech analysis have been used to establish the level of affinity of collaborative learning dialogues (Lubold & Pon-Harry, 2014), to evaluate the quality of oral presentations (Luzardo, Guamán, Chiluiza, Castells, & Ochoa, 2014), and to determine expertise in mathematics problem solving (Thompson, 2013).

Writing and Sketching

Two closely related modes are writing and sketching. They both use an instrument, most commonly a pen, to communicate complex thoughts. Using a pen is perhaps one of the first skills that students learn so using it to write and sketch is still a predominant activity in learning, especially at early stages. The most common information extracted from this mode is the transcript of what the student is saying, in the case of writing, or a structured representation of the sketches where information about their content can be inferred. However, capturing the process of writing and sketching through technological means opens the door to using information that human observers cannot easily detect, such as writing speed, rhythm, and pressure level. While their value for understanding learning is still not clear, there are indications that they could be good expertise predictors (Ochoa et al., 2013).

The recording instrument most commonly used to capture writing and sketching is a digital pen (Oviatt & Cohen, 2015). These pens are able to digitize the position, duration, and pressure of the strokes done on different surfaces. Once in digital form, this information can be used in LA tools. Alternatively, the widespread use of tablets in education (Clarke & Svanaes, 2014) also offers an opportunity to capture these modes easily, especially sketching.

In the realm of MLA, two works based on the math data corpus (Oviatt, Cohen, & Weibel, 2013) explored the contribution that writing and sketching modes could have in the prediction of expertise. Ochoa et al. (2013) extracted writing characteristics (stroke speed and length) and performed sketch recognition to determine the number of simple geometrical figures used. The results determined that speed of writing is highly correlated with level of expertise. Zhou, Hang, Oviatt, Yu, & Chen (2014) used classification systems based on writing and sketching characteristics to identify the expert in the group with 80% accuracy.

CONTEXTS

The main goal of MLA research is to extend the application of LA tools and methodologies to learning contexts that do not readily provide digital traces. One characteristic of these contexts is that the capture of more than one mode is necessary to understand the learning process. Table 11.1 presents a summary of the context studied in the current MLA literature with a detail of the modes used, the main learning aspects being explored in those contexts, and the works where those studies are conducted.

Table 11.1. Learning Contexts Studied by MLA

Contexts	Modes	Learning Aspects	Works
Lectures	Movement, Gaze, Gestures, Facial Expression, Speech	Attention, Question-Answer Interactions	Raca & Dillenbourg, 2013; Raca et al., 2014; Dominguez et al., 2015; D'Mello et al., 2015; Alibali et al., 2011
Oral Presentations	Posture, Movement, Gestures, Gaze, Speech, Digital Document	Skill Development, Feedback, Mental State	Luzardo et al., 2014; Echeverría et al., 2014; Chen et al., 2014; Leong et al., 2015; Schneider et al., 2015; Boncoddo et al., 2013
Problem-Solving	Movement, Actions, Speech, Writing, Sketching	Expertise Estimation	Ochoa et al., 2013; Luz, 2013; Thompson, 2013; Zhou et al., 2014
Construction Exercises	Gestures, Actions, Speech, Facial Expressions, Galvanic Skin Response	Novice vs. Expert Patterns	Worsley & Blikstein, 2013, 2014b, 2015a
Use of Intelligent Tutoring Systems	Digital Log Files, Facial Expressions, Speech	Relation between Emotions and Learning	Craig et al., 2008; D'Mello et al., 2008

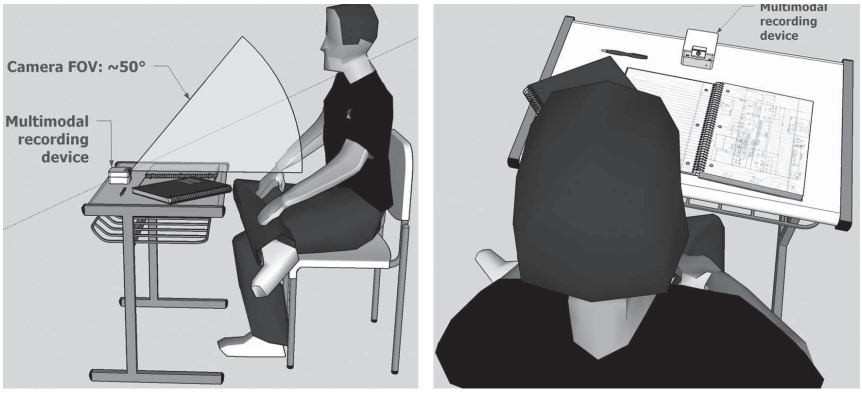


Figure 11.5. Design of a multimodal recording device (MRD) to be used in lecture settings. MRD in the classroom (left) and MRD from student's point of view (right)

Lectures

Traditionally, lectures are the most common context associated with learning. While several aspects of this setting deserve study, MLA researchers to date have focused on automatically assessing the attention of students during the lecture. The seminal works of Raca and Dillenbourg (2013) and Raca et al. (2014) have explored the recording of video in the classroom and the automatic extraction of student movement and gaze from that recording. The results of these studies suggest that both modes are related to student attention, but there are other significant contributors, such as sitting position. Dominguez, Echeverría, Chiluiza, and Ochoa (2015) presented a novel, distributed way to capture video-, audio- and pen-based modes using a multimodal recording device (MRD). Figure 11.5 presents the design of such a device. The proximity of the device to the students reduces the risk of occlusion and increases the video and audio capture quality. Finally, D'Mello et al. (2015) produced diverse audio recordings in a lecture setting in order to evaluate question–answer interactions between instructor and students.

Oral Presentations

The skill of presenting an academic topic in front of an audience is frequently regarded as one of the soft-skills that higher-education students should acquire (Debnath et al., 2012). Several independent groups around the globe have recently started to build MLA systems able to help novice students correct bad-practices and gain mastery in oral presentations. Echeverría et al. (2014) and Luzardo et al. (2014) present different aspects of the same system that uses gesture, posture, movement, gaze, speech, and an analysis of the digital presentation files and is able to predict the grade that a human evaluator will give the student. Chen, Leong, Feng, and Lee (2014), analyzing the same data, were able to combine the different modalities in composite variables also used to predict the score. Schneider, Börner, van Rosmalen, and Specht (2015) also created a virtual presentation skill trainer utilizing Kinect to recognize postures and provide feedback in real-time.

Problem-Solving

Learning, especially in STEM subjects, frequently occurs at individual and group problem-solving sessions (Silver, 2013). The existence of the math data corpus (Oviatt et al. 2013), a set of multimodal recordings of groups of three high-school students solving math and geometry problems, catalyzed MLA research in this setting. The media provided in the dataset include frontal video recordings of each student, video recordings of the working table, audio recordings of each student, and general audio of the room. Additionally, students were equipped with digital pens. Ground truth is provided about the level of expertise of the students. Luz (2013), Thompson (2013), Ochoa et al. (2013), and Zhou et al. (2014) have all analyzed this dataset using diverse modes, concluding that all the modalities contributed to the determination of the level of expertise with a high level of accuracy (>70%).

Construction Exercises

The knowledge and skills required for engineering design and construction can be tested through small construction challenges (Householder & Hailey, 2012). The seminal works of Worsley and Blikstein (2013, 2014b, 2015a) explore, through multimodal analysis, the patterns of actions performed by experts and novices in the design and manual assembly of structures. The main modes used for the analysis were gestures, actions, speech, facial expression, and galvanic skin response. The combination of traces extracted from these modes reveals differences in the construction process that are helpful to identify the level of mastery in engineering design.

Use of Intelligent Tutoring Systems

ITSs are usually studied by traditional LA using log-files. However, video and audio of the learner have

been captured to add new modes that complement the interaction data. The main modes extracted from the video are facial expression (Craig et al., 2008) and speech (D'Mello et al., 2008), which act as proxies for the learner's internal emotional state. Both are able to successfully detect emotional states such as boredom, confusion, and frustration in using the ITS.

SPECIFIC ISSUES

Once extracted, using multimodal traces in LA models and applications is similar to using different traces extracted from the same mode. However, MLA research and practice raise several specific issues when certain modalities are captured, processed, and analyzed. These issues remain open research areas, parallel to the technical extraction of traces from several modalities, but as important for the effective deployment of MLA solutions in the real-world.

Recording

Capturing interaction information in a digital tool is as easy and inexpensive as adding log statements in relevant parts of the code. These statements perform automatically, without requiring any involvement from the learner, in a transparent and generally reliable and error-free way. On the other hand, capturing media in the real-world requires the acquisition, installation, and use of recorders (cameras, microphones, digital pens, et cetera), turning the system on and off and monitoring it, and avoiding the degradation of the recording through occlusions, interference, or noise. Developing recording systems that work as effortlessly and efficiently as digital logging is one of the main barriers to the development of MLA. While this is an engineering problem, researchers should be aware of the feasibility and scalability of their solutions. One of the main proposals is to decentralize the recordings using inexpensive sensors that are always left on. If one or more recordings present problems, the general information could be reconstructed from the remaining working sensors.

Privacy

Capturing interaction information with digital tools already raises privacy concerns among students and instructors (Pardo & Siemens, 2014). The installation and use of recording systems that mimic "1984" levels of surveillance is bound to meet strong resistance. Informed consent forms could work for early research stages, but adopting MLA systems in the real-world would require a different, more creative approach. One of the most promising solutions in this area is transferring data ownership to the learner. Even if highly personal information is captured, privacy concerns are defused if the decision of what and when to share it remain in the control of the learner. This approach is similar to several quantified-self applications (Swan, 2013).

Integration

One question concerning the availability of large amounts of raw learning traces is how to combine them in order to produce useful information to understand and optimize the learning process. Traces extracted from different modes using different processes are bound to have very different characteristics. For example, the time granularity of the traces extracted from different modes can vary widely. Traces extracted from prosodic aspects of speech could change in tenths of a second while postures change more slowly. The level of certainty of the extracted traces can also be different. Speech recognition with high-quality recordings could reach 90% accuracy while emotional state detection from webcam sources could be in the low 70s. These difference do not prevent successful analysis, however, thoughtful design is required in order to prevent spurious results. Pioneering this line of research in MLA, Worsley and Blikstein (2014a) propose several fusion strategies based on the "bands of cognition" framework proposed by Newell (1994) and Anderson (2002) as an explanation for human cognition. The development of integration frameworks will benefit not only MLA but the whole LA community.

Impact on Learning

While the end-user tools and interventions based on multimodal learning analytics are similar to those based on monomodal analysis, the required usefulness of multimodal ones should be higher to justify the additional complexity of data acquisition. For example, a dashboard application based on data automatically captured by the LMS will be easier to accept than a similar dashboard that requires all classrooms be equipped with video cameras. The increased complexity should be accompanied by a larger positive impact on the learning process. The requirement of using multiple real-world signals to analyze learning should also come with the promise to provide more useful insights on the process and more measurable impacts on learners.

CONCLUSION

LA has revolutionized the approaches used to understand and optimize the learning process. However, its current bias towards studies and tools involving only computer-based learning contexts jeopardizes its applicability to learning in general. MLA is a subfield that seeks to integrate non-computer mediated learning contexts into the mainstream research and practice of LA.

This chapter presented the current-state-of-art in MLA. Modes as diverse as posture, speech, and sketching,

alongside the more traditional modes of clickstream information and textual content, has been used to answer research questions and to build feedback systems in learning contexts. A mixture of computer science techniques and insights provided by educational and behavioural scientists enable the automatic evaluation of very diverse learning contexts, such as classrooms, study groups, and oral presentations.

As can be inferred from the list of research presented in this chapter, MLA is still a nascent field with a small but very active and open community of researchers. The existence of regular challenges and workshops, where multimodal datasets are freely shared and jointly analyzed with new designs ideas openly discussed, creates a research environment where new knowledge is generated rapidly.

While several issues still prevent MLA from becoming a mainstream practice, active research projects are exploring solutions to those issues, making the capture of multimodal learning traces cheaper, less invasive, and more automatic. Novel solutions born from the MLA community to handle privacy concerns, such as providing distributed recording and resting the ownership of the data with the learner, could one day be the norm for general LA practices.

Finally, the author wishes to invite LA researchers and practitioners to explore the use of multiple modalities in their own studies and tools. The MLA community will openly share its knowledge, data, code, and frameworks. Only the embrace of these different modalities will allow LA to have an impact in all the contexts where learning takes place.

REFERENCES

Alibali, M. W., Nathan, M. J., Fujimori, Y., Stein, N., & Raudenbush, S. (2011). Gestures in the mathematics classroom: What's the point? In N. Stein & S. Raudenbush (Eds.), *Developmental cognitive science goes to school* (pp. 219–234). New York: Routledge, Taylor & Francis.

Almeda, M. V., Scupelli, P., Baker, R. S., Weber, M., & Fisher, A. (2014). Clustering of design decisions in classroom visual displays. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 44–48). New York: ACM.

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science*, *26*(1), 85–112.

Arnold, K. E., & Pistilli, M. D. (2012). Course Signals at Purdue: Using learning analytics to increase student success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge* (LAK '12), 29 April–2 May 2012, Vancouver, BC, Canada (pp. 267–270). New York: ACM.

Blikstein, P. (2013). Multimodal learning analytics. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 102–106). New York: ACM.

Boraston, Z., & Blakemore, S.-J. (2007). The application of eye-tracking technology in the study of autism. *The Journal of Physiology*, *581*(3), 893–898.

Bull, P. E. (2013). *Posture & Gesture*. Elsevier.

Boncoddo, R., Williams, C., Pier, E., Walkington, C., Alibali, M., Nathan, M., Dogan, M. & Waala, J. (2013). Gesture as a window to justification and proof. In M. V. Martinez & A. C. Superfine (Eds.), *Proceedings of the 35th annual meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (PME-NA 35) 14–17 November 2013, Chicago, IL, USA (pp. 229–236). http://www.pmena.org/proceedings/

Chen, L., Leong, C. W., Feng, G., & Lee, C. M. (2014). Using multimodal cues to analyze MLA '14 oral presentation quality corpus: Presentation delivery and slides quality. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 45–52). New York: ACM.

Clarke, B., & Svanaes, S. (2014, April 9). An updated literature review on the use of tablets in education. Family Kids and Youth. https://smartfuse.s3.amazonaws.com/mysandstorm.org/uploads/2014/05/T4S-Use-of-Tablets-in-Education.pdf

Cobb, P., Confrey, J., Lehrer, R., Schauble, L., & others. (2003). Design experiments in educational research. *Educational Researcher*, *32*(1), 9–13.

Craig, S. D., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the facial action coding system to cognitive–affective states during learning. *Cognition and Emotion*, *22*(5), 777–788.

Crossley, S. A., Roscoe, R. D., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. *Proceedings of the 26th Annual Florida Artificial Intelligence Research Society Conference* (FLAIRS-13), 20–22 May 2013, St. Pete Beach, FL, USA (pp. 208–213). Menlo Park, CA: The AAAI Press.

Debnath, M., Pandey, M., Chaplot, N., Gottimukkula, M. R., Tiwari, P. K., & Gupta, S. N. (2012). Role of soft skills in engineering education: Students' perceptions and feedback. In C. S. Nair, A. Patil, & P. Mertova (Eds.), *Enhancing learning and teaching through student feedback in engineering* (pp. 61–82). ScienceDirect. http://www.sciencedirect.com/science/book/9781843346456

D'Mello, S. K., Jackson, G. T., Craig, S. D., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., el Kaliouby, R., Picard, R., & Graesser, A. C. (2008). AutoTutor detects and responds to learners' affective and cognitive states. Workshop on Emotional and Cognitive Issues in ITS, held in conjunction with the 9th International Conference on Intelligent Tutoring Systems (ITS 2008), 23–27 June 2008, Montreal, PQ, Canada. https://www.researchgate.net/publication/228673992_AutoTutor_detects_and_responds_to_learners_affective_and_cognitive_states

D'Mello, S., Olney, A., Blanchard, N., Samei, B., Sun, X., Ward, B., & Kelly, S. (2015). Multimodal capture of teacher–student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 17th ACM International Conference on Multimodal Interaction* (ICMI '15), 9–13 November 2015, Seattle, WA, USA (pp. 557–566). New York: ACM.

Dominguez, F., Echeverría, V., Chiluiza, K., & Ochoa, X. (2015). Multimodal selfies: Designing a multimodal recording device for students in traditional classrooms. *Proceedings of the 17th ACM International Conference on Multimodal Interaction* (ICMI '15), 9–13 November 2015, Seattle, WA, USA (pp. 567–574). New York: ACM.

Echeverría, V., Avendaño, A., Chiluiza, K., Vásquez, A., & Ochoa, X. (2014). Presentation skills estimation based on video and Kinect data analysis. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 53–60). New York: ACM.

Freedman, D. H. (2010, December 10). Why scientific studies are so often wrong: The streetlight effect. *Discover Magazine*, *26*. http://discovermagazine.com/2010/jul-aug/29-why-scientific-studies-often-wrong-streetlight-effect

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694–724.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction*. Longman Publishing.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612–618.

Householder, D. L., & Hailey, C. E. (2012). *Incorporating engineering design challenges into STEM courses*. National Center for Engineering and Technology Education. http://ncete.org/flash/pdfs/NCETECaucusReport.pdf

Jewitt, C. (2006). *Technology, literacy and learning: A multimodal approach*. Psychology Press.

Kizilcec, R. F., Piech, C., & Schneider, E. (2013). Deconstructing disengagement: Analyzing learner subpopulations in massive open online courses. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 170–179). New York: ACM.

Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Edward Arnold.

Krugman, D. M., Fox, R. J., Fletcher, J. E., Fischer, P. M., & Rojas, T. H. (1994). Do adolescents attend to warnings in cigarette advertising? An eye-tracking approach. *Journal of Advertising Research*, *34*, 39–51.

Kruschke, J. K. (2003). Attention in learning. *Current Directions in Psychological Science*, *12*(5), 171–175.

Leong, C. W., Chen, L., Feng, G., Lee, C. M., & Mulholland, M. (2015). Utilizing depth sensors for analyzing multimodal presentations: Hardware, software and toolkits. *Proceedings of the 17th ACM International Conference on Multimodal Interaction* (ICMI '15), 9–13 November 2015, Seattle, WA, USA (pp. 547–556). New York: ACM.

Lin, Y.-T., Lin, R.-Y., Lin, Y.-C., & Lee, G. C. (2013). Real-time eye-gaze estimation using a low-resolution webcam. *Multimedia Tools and Applications*, *65*(3), 543–568.

Lubold, N., & Pon-Barry, H. (2014). Acoustic-prosodic entrainment and rapport in collaborative learning dialogues. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 5–12). New York: ACM.

Lund, K. (2007). The importance of gaze and gesture in interactive multimodal explanation. *Language Resources and Evaluation*, *41*(3–4), 289–303.

Luzardo, G., Guamán, B., Chiluiza, K., Castells, J., & Ochoa, X. (2014). Estimation of presentations skills based on slides and audio features. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 37–44). New York: ACM.

Luz, S. (2013). Automatic identification of experts and performance prediction in the multimodal math data corpus through analysis of speech interaction. *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI '13), 9–13 December 2013, Sydney, Australia (pp. 575–582). New York: ACM.

Markaki, V., Lund, K., & Sanchez, E. (2015). Design digital epistemic games: A longitudinal multimodal analysis. Paper presented at the conference *Revisiting Participation: Language and Bodies in Interaction*, 24–27 June 2015, Basel, Switzerland.

Mazur-Palandre, A., Colletta, J. M., & Lund, K. (2014). Context sensitive "how" explanation in children's multimodal behavior, *Journal of Multimodal Communication Studies*, 2, 1–17.

Mishra, B., Fernandes, S. L., Abhishek, K., Alva, A., Shetty, C., Ajila, C. V., … Shetty, P. (2015). Facial expression recognition using feature based techniques and model based techniques: A survey. *Proceedings of the 2nd International Conference on Electronics and Communication Systems* (ICECS 2015), 26–27 February 2015, Coimbatore, India (pp. 589–594). IEEE.

Mitra, S., & Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *37*(3), 311–324.

Morency, L.-P., Oviatt, S., Scherer, S., Weibel, N., & Worsley, M. (2013). ICMI 2013 grand challenge workshop on multimodal learning analytics. *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI '13), 9–13 December 2013, Sydney, Australia (pp. 373–378). New York: ACM.

Newell, A. (1994). *Unified theories of cognition*. Harvard University Press.

Ochoa, X., Chiluiza, K., Méndez, G., Luzardo, G., Guamán, B., & Castells, J. (2013). Expertise estimation based on simple multimodal features. *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI '13), 9–13 December 2013, Sydney, Australia (pp. 583–590). New York: ACM.

Ochoa, X., Worsley, M., Chiluiza, K., & Luz, S. (2014). MLA '14: Third multimodal learning analytics workshop and grand challenges. *Proceedings of the 16th ACM International Conference on Multimodal Interaction* (ICMI '14), 12–16 November 2014, Istanbul, Turkey (pp. 531–532). New York: ACM.

Oviatt, S., Cohen, A., & Weibel, N. (2013). Multimodal learning analytics: Description of math data corpus for ICMI grand challenge workshop. *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI '13), 9–13 December 2013, Sydney, Australia (pp. 583–590). New York: ACM.

Oviatt, S., & Cohen, P. R. (2015). *The paradigm shift to multimodality in contemporary computer interfaces*. San Rafael, CA: Morgan & Claypool Publishers.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research. *Encyclopedia of Human Computer Interaction*, *1*, 211–219.

Raca, M., & Dillenbourg, P. (2013). System for assessing classroom attention. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 265–269). New York: ACM.

Raca, M., Tormey, R., & Dillenbourg, P. (2014). Sleepers' lag: Study on motion and attention. *Proceedings of the 4th International Conference on Learning Analytics and Knowledge* (LAK '14), 24–28 March 2014, Indianapolis, IN, USA (pp. 36–43). New York: ACM.

Scherer, S., Worsley, M., & Morency, L.-P. (2012). 1st international workshop on multimodal learning analytics. *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (ICMI '12), 22–26 October 2012, Santa Monica, CA, USA (pp. 609–610). New York: ACM.

Schlömer, T., Poppinga, B., Henze, N., & Boll, S. (2008). Gesture recognition with a Wii controller. *Proceedings of the 2nd International Conference on Tangible and Embedded Interaction* (TEI '08), 18–21 February 2008, Bonn, Germany (pp. 11–14). New York: ACM.

Schneider, J., Börner, D., van Rosmalen, P., & Specht, M. (2015). Presentation trainer, your public speaking multimodal coach. *Proceedings of the 17th ACM International Conference on Multimodal Interaction* (ICMI '15), 9–13 November 2015, Seattle, WA, USA (pp. 539–546). New York: ACM.

Serrano-Laguna, A., & Fernández-Manjón, B. (2014). Applying learning analytics to simplify serious games deployment in the classroom. *Proceedings of the 2014 IEEE Global Engineering Education Conference* (EDUCON 2014), 3–5 April 2014, Istanbul, Turkey (pp. 872–877). IEEE.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 9, 51–60. doi:0002764213498851.

Silver, E. A. (2013). *Teaching and learning mathematical problem solving: Multiple research perspectives*. Routledge.

Simsek, D., Sándor, Á., Buckingham Shum, S., Ferguson, R., De Liddo, A., & Whitelock, D. (2015). Correlations between automated rhetorical analysis and tutors' grades on student essays. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 355–359). New York: ACM.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big Data*, 1(2), 85–99.

Thompson, K. (2013). Using micro-patterns of speech to predict the correctness of answers to mathematics problems: An exercise in multimodal learning analytics. *Proceedings of the 15th ACM International Conference on Multimodal Interaction* (ICMI '13), 9–13 December 2013, Sydney, Australia (pp. 591–598). New York: ACM.

Vasquez, H. A., Vargas, H. S., & Sucar, L. E. (2015). Using gestures to interact with a service robot using Kinect 2. *Advances in Computer Vision and Pattern Recognition*, 85. Springer.

Vatrapu, R., Reimann, P., Bull, S., & Johnson, M. (2013). An eye-tracking study of notational, informational, and emotional aspects of learning analytics representations. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 125–134). New York: ACM.

Worsley, M., & Blikstein, P. (2013). Towards the development of multimodal action based assessment. *Proceedings of the 3rd International Conference on Learning Analytics and Knowledge* (LAK '13), 8–12 April 2013, Leuven, Belgium (pp. 94–101). New York: ACM.

Worsley, M., & Blikstein, P. (2014a). Deciphering the practices and affordances of different reasoning strategies through multimodal learning analytics. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 21–27). New York: ACM.

Worsley, M., & Blikstein, P. (2014b). Using multimodal learning analytics to study learning mechanisms. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the 7th International Conference on Educational Data Mining* (EDM2014), 4–7 July, London, UK (pp. 431–432). International Educational Data Mining Society.

Worsley, M., & Blikstein, P. (2015a). Leveraging multimodal learning analytics to differentiate student learning strategies. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 360–367). New York: ACM.

Worsley, M., & Blikstein, P. (2015b). Using learning analytics to study cognitive disequilibrium in a complex learning environment. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 426–427). New York: ACM.

Zhang, Z. (2012). Microsoft Kinect sensor and its effect. *IEEE MultiMedia*, 19(2), 4–10.

Zhou, J., Hang, K., Oviatt, S., Yu, K., & Chen, F. (2014). Combining empirical and machine learning techniques to predict math expertise using pen signal features. *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge* (MLA '14), 12–16 November 2014, Istanbul, Turkey (pp. 29–36). New York: ACM.