Multimodal Learning Analytics as a Tool for Bridging Learning Theory and Complex Learning Behaviors

Marcelo Worsley
Stanford University
520 Galvez Mall, CERAS 217
Stanford, CA 94305
mworsley@stanford.edu

ABSTRACT

The recent emergence of several low-cost, high resolution, multimodal sensors has greatly facilitated the ability for researchers to capture a wealth of data across a variety of contexts. Over the past few years, this multimodal technology has begun to receive greater attention within the learning community. Specifically, the Multimodal Learning Analytics community has been capitalizing on new sensor technology, as well as the expansion of tools for supporting computational analysis, in order to better understand and improve student learning in complex learning environments. However, even as the data collection and analysis tools have greatly eased the process, there remain a number of considerations and challenges in framing research in such a way that it lends to the development of learning theory. Moreover, there are a multitude of approaches that can be used for integrating multimodal data, and each approach has different assumptions and implications. In this paper, I describe three different types of multimodal analyses, and discuss how decisions about data integration and fusion have a significant impact on how the research relates to learning theories.

Categories and Subject Descriptors

K.3.m [**Computers and Education**]: Computer Uses in Education – M*iscellaneous.*

Keywords

Learning Sciences; Constructionism; Cognition

1. INTRODUCTION

Over the past few years multimodal learning analytics [1 - 4] has become an increasingly prevalent paradigm for studying and improving complex learning environments. However, even as the field has started to develop an identity, there remains a wide range of research that gets cast under the guise of multimodal learning analytics. In this paper, I provide a framework and a set of terminology that can be used to both characterize and advance the types of multimodal learning analytics research that we, as a field, pursue. Specifically, I describe three studies that represent different approaches for studying complex learning environments through multimodal learning analytic techniques. Each approach represents a different underlying frame for how multimodal data streams are fused. The three approaches that I discuss are the naïve fusion frame, the low-level fusion frame, and high-level fusion frame. These three frames are not expected to encompass all research that would be categorized under the heading of multimodal learning analytics, but likely represent the simplest, most common, and perhaps, most important analytic approaches.

In the sections to follow, I present an analysis from the perspective of each of the aforementioned frames, and then discuss some of the affordances and drawbacks that they confer. However, before describing each frame in detail, I highlight prior research that will be important for the forthcoming discussion of the three different approaches.

2. PRIOR LITERATURE

This paper builds on [5], which describes various “bands of cognition,” and later work by [6]. Specifically, [5] describes time scales across which human actions can be interpreted as biological, cognitive, rational and social. Each band captures three orders of magnitude beginning from 100s of microseconds (10^{-4} seconds), all the way up to months (10^{7} seconds). Specifically, the biological band is centered on time scales of a microsecond; the cognitive band on time scales of seconds; the rational band on time scales of 10 minutes; and the social band on time scales of weeks. The framework also describes how each time scale is associated with different levels of intentionality, and different types of activities. For example, actions that take place within the biological band are sometimes interpreted as occurring at an unconscious, non-deliberate level, whereas, completion of a task is normally associated with human actions in the rational band. [6] builds on this framework by considering the extent to which human actions that occur within a very short time scale, i.e. one of the lower bands (biological band or cognitive band), influence human actions at larger time scales, i.e. the rational band and the social Band. In discussing a bridge across time spans, [6] proposes three theses: the Decomposition Thesis, the Relevance Thesis and the Modeling Thesis. The Decomposition Thesis claims that the events that occur at longer time scales, can be decomposed into actions on shorter time scales. The Relevance Thesis relates to the claim that the “microstructure of cognition is relevant for educational issues.” In practical terms, this means that short time scale actions are important for studying and diagnosing learning development. Finally the Modeling Thesis is concerned with the ability for cognitive modeling to help explain how to use the fine-grained information to improve instruction.

In the context of the bands of cognition and the various theses proposed by [6], this paper can be seen as describing ways that multimodal learning analytics has relevance at different time scales, and their associated bands of cognition. Specifically, when fusing different data streams, an important consideration will be the time scale(s) that are being used, and the time scale(s) of the results that are presented. Decisions about each of these will be central to the analysis’ utility for relating to, or building on, learning theory. These decisions will also have a significant impact on the implications derived from the analysis.

MLA'14, November 12-16 2014, Istanbul, Turkey
Copyright 2014 ACM 978-1-4503-0488-7/14/11$15.00
http://dx.doi.org/10.1145/2666633.2666634.

3. NAÏVE FUSION/CLASSIFIER FRAME

I begin with the Naïve Classifier Frame because it, in many respects, represents the simplest approach used in conducting multimodal learning analytics research. This particular approach is typified by the integration of aggregate features from a variety of modalities, without a specific hypothesis or set of assumptions about how those features interact with one another - this is the basis for using the term 'naïve.' At the same time it is often the approach used to conduct exploratory research. That the approach is termed 'naïve' should not, however, be taken to mean that the features used from each of the different modalities are without theoretical merit. Instead, researchers often use prior experience, and prior literature in order to inform which features they will consider in their analysis. To make this clearer, I use an example from my prior work that examines expertise in an engineering design context [7].

The data analyzed was derived from eighteen students at a tier-1 research university. This population of students included everything from undergraduate humanities majors, to PhD level engineering graduate students. Student's prior experience was used to label them as either a novice, an intermediate or an expert. During the study, each student was asked to individually design an automatic trash separation system that could distinguish between glass, paper, plastic and metal. As students engaged in a think-aloud protocol, I collected, audio and video data, in addition to their design drawings.

For this specific study, I was able to conduct various individual analyses based on: content word from science, technology, engineering and mathematics (STEM) domains; speech (prosodic and spectral features), dependency parsing; scientific argumentation; sentiment; and drawing. These features were selected based on prior work that found correlations between expertise and scientific argumentation [8], sentiment [9 - 11], language and speech [12, 13], uncertainty [13], to name a few. Furthermore, the design of the study was informed by previous work from the learning sciences that uses interviews to study experts and novices (e.g. [14 – 16]). However, when conducting the analysis I did not have a specific theoretical framework for describing how the different modalities interacted with one another. Hence, my approach was to use natural demarcations in the interview and examine aggregate summary statistics, i.e. minimum, maximum and mean, for each modality.

Using the data extracted from the different techniques, I used a combination of feature reduction algorithms to pinpoint the features that (1) most closely aligned to student expertise and (2) that seemed most appropriate for inclusion. Some of these features included student certainty, sentiment, adaptive tool usage and the frequencies of strategic and schematic utterances. Thus, simply completing feature reduction proved to be a useful entry point for identifying practices and behaviors associated with expertise.

After identifying the appropriate features from each modality, I used those features to train an unsupervised model that predicted expertise among the three possible class labels, at 87% accuracy, which was significant given than humans achieved less than 50% accuracy, when judging participant expertise based on transcripts of each participant's response.

Utilizing a process that involves identification of aggregate features that correlate with one's dependent variable is a good starting point for exploratory analysis of student behaviors. Those features can subsequently be used to create a model or classifier and iteratively improved based on one's objective. However, basing one's analysis on aggregate features may overlook some of the nuances of the data, especially in the case where the researcher has a specific question in mind. Moreover, in the context of Newell's bands of cognition, taking aggregate measures from an entire experiment may only be useful for identifying features that exist in the rational or social bands.

4. LOW-LEVEL FUSION FRAME

The second approach that I present is one that fits most models of low-level fusion. In this paradigm, the researcher is intentional about enacting multimodal data fusion on very small time scales. One reason for doing this is because the researcher may have prior knowledge that the various modalities have time specific relevance to one another. For example, the participant's average pitch may be of less importance, than their pitch in the context of their immediate actions, or gestures. To better describe this approach, I will again present an example from my own work [17].

The particular instance that I describe, builds on work from epistemological frames [18, 19]. The work of [18] identifies student epistemological frames based on multimodal behaviors. Specifically, they found that a combination of posture, gaze, gesturing and speech, could be used to typify four very distinct epistemological frames that students use in a multi-person, collaborative problem-solving setting. Research by [19] expands on this work by describing the epistemological frames that students use during informal cognitive clinical interviews. Within these interviews, the authors again used multimodal data in order to describe the characteristic behaviors of an expert frame, an inquiry frame and an oral examination frame.

Accordingly, my use of multimodal analysis in the low-level fusion frame, is motivated be a desire to capture similar epistemological frames as previous authors. However, my context is somewhat different, in that I am studying pairs of students as they complete an engineering design task. Because of this, I do not assume that the common multimodal behaviors will be the same. To address this, and identify what the common frames in the context of my study, I collect audio, hand/wrist movement and electro-dermal activation data. I fuse the three data streams, on a per second basis, between each pair of design trial that the students attempt. For a given second in time a participant can be described based on whether or not they generated audio, their average hand/wrist displacement, and their average electro-dermal activation value. Note: data from each sensor was captured at a different time scale, thus it was advantageous to find a lowest common denominator and use that for data fusion.

Those values were then used to populate a matrix that included all data points for all students. Each column of the matrix was normalized, and then processed through X-Means clustering to identify the common behaviors. Figure 1 contains the cluster centroids for the four common behaviors that emerged.

From the cluster centroids there is a cluster typified by high amounts of audio, another typified by high amounts of stress, and another typified by high amounts of hand/wrist movement. Interestingly, though, the fourth cluster is defined as having average values across all three dimensions, which I have entitled FLOW. What is not reflected in this figure though, is that the vast majority of user actions fall into the FLOW behavior Figure 2.

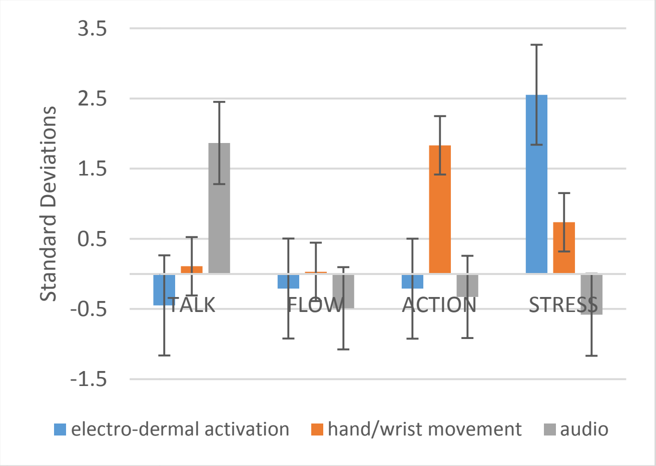

Figure 1. Cluster centroids for common multimodal behaviors.

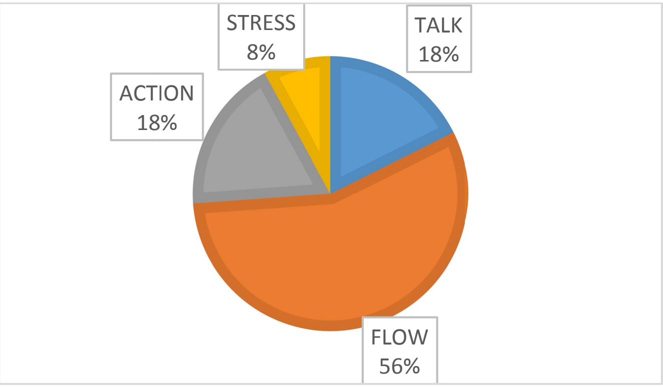

Figure 2. Overall proportion of cluster use.

The importance of the FLOW behavior becomes increasingly significant when considering that this particular study involved two different experimental conditions, and that the two experimental conditions significantly differed in their usage of FLOW. Additionally, when I compare the participants of two conditions in the overall similarity of their sequence of behaviors, I also see statistically significant differences. In this way, taking a more targeted, and theoretically driven approach in the fusion of the multimodal data provided significant benefit in highlighting the differences between the two conditions. Furthermore, taking this approach provided a means for recognizing differences at multiple time scales, and at different bands of cognition.

5. HIGH-LEVEL FUSION FRAME

This final approach is very similar to that of low-level fusion, but differs in that the multimodal data fusion happens at a different scale. Instead of merging raw data, one or more modalities, first undergoes semantic analysis that attempts to make sense of that raw data. As I will describe in the following example, in lieu of using raw hand/wrist data, I instead classify human hand/wrist movements into one of six possible gestures (Table 1).

In this way, then, the assumption is that speech and/or electro-dermal activation have more to do with specific *types* of hand/wrist movements, as opposed to having relevance to specific *amounts* of hand/wrist displacement. One of the theoretical underpinnings for this type of analysis is [20] which involves examining teacher and student gesturing in the context of language.

Table 1. Object Manipulation Classes

Class	Codes
PLAN	Prototyping ideas or inspecting the materials
EVALUATE	Testing a mechanism or testing the system
MODIFY	Making changes to an existing design
NOTHING	Not actively engaging in the activity
REVERT	Undoing one of more parts of a previous design
REALIZE	Putting pieces together as to make the structure

Accordingly, they are interested in when individuals use speech, for example, in the context of a specific type of gesture (e.g. pointing or underlining). Simply measuring the amount of hand/wrist displacement in that context would likely be of little importance, since many gestures may be ambiguous when classified based on displacement.

The example analysis that I present uses the same data as the low-level fusion frame. However, as noted, instead of including the hand/wrist displacement, I included a five dimensional vector that includes a semantic interpretation of what the student was hoping to accomplish with their hand/wrist movements. Apart from that difference, the analysis proceeds as beforehand, by starting with the identification of common multimodal behaviors. Unlike low-level fusion, this particular approach did not yield significant findings in terms of the behavioral differences that distinguished one experimental condition from the other. Instead, this particular approach proved to be most effective at providing insights into how students from the different experimental conditions enacted the Object Manipulation Classes, or semantic hand/wrist movements, differently. It also provided a means for predicting when students when students would be unable to complete the activity, and when student learning was unlikely. Accordingly, there were important results to be revealed, but one take away is that moving to higher level interpretations of user actions may mask certain behavioral differences, or may require that researchers also use high level data from other modalities as well. Furthermore, while the specific algorithm used in the high-level fusion was the same as that from low-level fusion, changing the way that data fusion took place greatly impacted the type of information that the analysis provided.

6. DISCUSSION

The three data fusion approaches presented in the previous sections have utility for different types of analyses. Naïve fusion was useful for conducting exploratory work on the features that are salient to one's dependent variable, and could do so at the level of participants' summary statistics. This has clear relevance as an entry point into unraveling multimodal data, as well as for building classifiers. However, the cautious nature of this approach makes it less effective in considering the temporal aspects of each participant's process. Additionally, the statistical significance of one's results can be greatly weakened by such exploratory work especially, when properly taking into account multiple comparison testing and post-hoc analysis.

Low-level fusion, on the other hand, provided a means to more closely consider the temporal elements of each participant's process, and, specifically, included important information about the context in which each data point emerged (relative to data from the other modalities). For analyses that want to deeply draw upon context across modalities, taking the low-level fusion approach can be quite informative, and, based on the example presented in this paper, provide insights that span multiple time scales.

Lastly, high-level fusion as presented here tries to draw additional semantic from the raw, low-level data, such that seemingly similar patterns at the low-level can be properly binned based on the user's intentions, for example. This is important when the low-level data does not properly capture the specific level of analysis, learning theory, or band of cognition that the researcher is interested in. However, as observed in the example, one cannot always expect for low-level fusion, and high-level fusion to provide the same results. Instead, based on prior work from psychology, one could actually expect for these to be quite different. That said, high-level fusion has significant potential to provide deep insights into how the semantic actions or behaviors are being enacted.

7. CONCLUSION

With the advent of new multimodal sensor technology, multimodal learning analytics appears to be getting easier. However, as one is embarking on this form of analysis, it is important to carefully consider the theoretical frame being used, as this may have a large consequences on the results one attains. In this short paper I have presented examples from three different data fusion strategies, all of which have different affordances and different drawbacks. As previously noted, this list of fusion strategies is not exhaustive, and does not begin to consider the complexities introduced by considering that direct time alignment is not necessarily appropriate for various types of data fusion.

While the goal of this paper has not been to suggest that any one approach is better than the other, I do generally advocate for techniques that have a clear connection to previous learning theory, or that are in pursuit of new learning theory. Additionally, I think it is important to consider how one's analysis can relates to learning practices and behaviors. That said, there are opportunities to do this using any of the three data fusion techniques, the challenge, however, is determining the level of analysis at which your studies have relevance and to think about how to use those finding to make bridges to learning pedagogy.

8. REFERENCES

[1] Worsley, M. 2012. Multimodal Learning Analytics: Enabling the Future of Learning through Multimodal Data Analysis and Interfaces. In Proceedings of the 14th ACM international conference on Multimodal Interaction (ICMI '12). ACM, New York, NY, USA. 353-356.

[2] Blikstein, P. 2013. Multimodal Learning Analytics. In Proceedings of the 3rd Annual Learning Analytics and Knowledge Conference. Leuven, Belgium.

[3] Scherer,S., Worsley,M. and Morency, L. 2012. 1st international workshop on multimodal learning analytics: extended abstract. In Proceedings of the 14th ACM international conference on Multimodal interaction (ICMI '12). ACM, New York, NY, USA, 609-610.

[4] Morency, L. Oviatt, S., Scherer, S. Weibel, N. and Worsley, M. (2013). ICMI 2013 Grand Challenge Workshop on Multimodal Learning Analytics. In Proceedings of the 15th ACM International Conference on Multimodal interaction (ICMI '13). ACM, New York, NY, USA,

[5] Newell, A. 1994. Unified theories of cognition. Harvard University Press.

[6] Anderson, J. 2002. Spanning seven orders of magnitude: a challenge for cognitive modeling. Cognitive Science, 26(1), 85–112.

[7] Worsley, M., & Blikstein, P. 2011. What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis. In Proceedings of the Fourth Annual Conference on Educational Data Mining (EDM 2011) (pp. 235–240). Eindhoven, Netherlands.

[8] Li, M., Ruiz-Primo, M. A., & Shavelson, R. J. 2006. Towards a science achievement framework: The case of TIMSS 1999. In S. Howie & T. Plomp (Eds.), Contexts of learning mathematics and science: Lessons learned from TIMSS (pp. 291–311). London: Routledge.

[9] Craig, S. D., D'Mello,S., Witherspoon, A. and Graesser, A. 2008. 'Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive-affective states during learning', Cognition & Emotion, 22: 5, 777 — 788.

[10] D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., and Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. User Modeling and User-Adapted Interaction 18, 1-2 (Feb. 2008), 45-80.

[11] Conati, C. and Maclaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction. August, 2009 267-303.

[12] Litman, D., Moore, J., Dzikoska, M., and Farrow. E. 2009. Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains and Modalities. Proceedings 14th International Conference on Artificial Intelligence in Education (AIED), Brighton, UK, July.

[13] Liscombe, J., Hisrchberg, J., and Venditti, J. 2005. Detecting Certainness in Spoken Tutorial Dialogues. In Proceedings of Interspeech 2005—Eurospeech, Lisbon, Portugal.

[14] Bransford, J., Brown, A., & Cocking, R. 2000. How people learn.

[15] Chi, M. Glaser, Rees. 1981. Expertise in problem solving.

[16] Ahmed, S., & Wallace, K. M. 2003. Understanding the differences between how novice and experienced designers approach design tasks, 14, 1–11.

[17] Worsley, M. Blikstein, P. (in press). Deciphering the Practices and Affordances of Different Reasoning Strategies through Multimodal Learning Analytics. In *Proceedings of the 2014 Multimodal Learning Analytics Workshop and Grand Challenge*.

[18] Scherr, R. E., & Hammer, D. 2009. Student Behavior and Epistemological Framing: Examples from Collaborative Active-Learning Activities in Physics. Cognition and Instruction, 27(2), 147–174.

[19] Russ, R. S., Lee, V. R., & Sherin, B. L. (2012). Framing in cognitive clinical interviews about intuitive science knowledge: Dynamic student understandings of the discourse interaction. Science Education, 96(4), 573–599.

[20] Alibali, M. W., & Nathan, M. J. (2012). Journal of the Learning Embodiment in Mathematics Teaching and Learning : Evidence From Learners ' and Teachers ' Gestures, 37–41.