Exploring Behavior Representation for Learning Analytics

Marcelo Worsley
Rossier School of Education
& Institute for Creative
Technologies University of
Southern California
Los Angeles, CA, USA
worsley@usc.edu

Stefan Scherer
Institute for Creative
University of Southern California
Playa Vista, CA, USA
scherer@ict.usc.edu

Louis-Philippe Morency
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA
morency@cs.cmu.edu

Paulo Blikstein
Graduate School of Education
Stanford University
Stanford, CA, USA
paulob@stanford.edu

ABSTRACT
Multimodal analysis has long been an integral part of studying learning. Historically multimodal analyses of learning have been extremely laborious and time intensive. However, researchers have recently been exploring ways to use multimodal computational analysis in the service of studying how people learn in complex learning environments. In an effort to advance this research agenda, we present a comparative analysis of four different data segmentation techniques. In particular, we propose affect- and pose-based data segmentation, as alternatives to human-based segmentation, and fixed-window segmentation. In a study of ten dyads working on an open-ended engineering design task, we find that affect- and pose-based segmentation are more effective, than traditional approaches, for drawing correlations between learning-relevant constructs, and multimodal behaviors. We also find that pose-based segmentation outperforms the two more traditional segmentation strategies for predicting student success on the hands-on task. In this paper we discuss the algorithms used, our results, and the implications that this work may have in non-education-related contexts.

Categories and Subject Descriptors
H.5.m **[Information Interfaces and Presentation]**: Miscellaneous
J.4 **[Computer Applications]**: Social and Behavioral Sciences

General Terms
Algorithms, Human Factors, Experimentation, Performance.

Keywords
Learning Sciences, Interaction Analysis, Modeling

1. INTRODUCTION
Multimodal learning analytics [3,22,28] is still a nascent field. Over the past few years there has been increasing interest and participation in this research domain [11,12,18]. As researchers move towards integrating multimodal human perception with education research, there are a host of important questions that need to be explored.

As in all studies that leverage multimodal data, there is the question of how to appropriately represent the student's behaviors. This entails selecting a set of modalities which reliably capture behaviors that correlate with the dependent variable. As an additional question, there are a host of constructs - success, persistence, learning, motivation – that may all be of interest for a given learning analytics study. Once the dependent variable and the modalities have been selected, the researcher must then wrestle with how to utilize the data to map between user behavior and the dependent variable. One element of mapping between data and a learning construct is closely linked to how the data is segmented. Comparing different segmentation methods is a primary contribution of this paper.

In multimodal learning analytics research, and multimodal interaction research more generally, researchers often use one of two approaches for segmenting data. Some rely on human-annotation to properly denote an appropriate segment of interaction. This approach is extremely laborious, but has the advantage of incorporating elements of human perception that can enable detection of hard to identify interactions of verbal and non-verbal cues. Another advantage is that human-based segmentation can handle variable length segments. Our prior research in multimodal learning analytics [24,25] and that of many others, was based on human segmentation. At the other end of the spectrum is fixed-, or sliding-, window segmentation. Researchers select the window size that garners the best results, or that was established by prior research, to automatically segment the data. This approach has the advantage of being far less laborious, but assumes that the appropriate unit of analysis is fixed in length, and is tied to a certain amount of time. In this paper we propose two approaches that look to simultaneously harness the affordances of human-based segmentation and fixed-window segmentation. In particular, we compare affect-based segmentation, and pose-based segmentation to more traditional models of human-based segmentation and fixed-window segmentation, for modeling student behavior and learning. These different approaches will be compared to see how well they can be used to model and predict student behaviors as it relates to three constructs: success, experimental condition and learning. (Each of these constructs will be described in more details in the methods section.) (Note: some researchers simplify the process of

ICMI '15, November 09-13, 2015, Seattle, WA, USA
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11…$15.00
DOI: http://dx.doi.org/10.1145/2818346.2820737

human-based segmentation by incorporating ways that allow the data to be collected on a per problem, per task or per session basis, for example. This often times provides a representation that is too coarse, but can offer a middle ground between human annotation and automated techniques).

Our hypothesis is that affect- and pose-based segmentation will offer a means for providing a semantically relevant, automated approach for understanding and predicting learning-relevant constructs that is as good as, or better than, human-based segmentation and/or fixed-window segmentation.

In what follows we briefly present our theoretical framework and pertinent prior literature; describe the experiment from which the data was derived; delineate the basic algorithm used to analyze the data; summarize important results; and discuss the implications of this work.

2. PRIOR LITERATURE

2.1 Theoretical Framework

The current analyses are influenced by prior literature on Interaction Analysis [13], Embodied Cognition [16], Epistemological Framing [20,23] and Cognitive Disequilibrium [8].

In particular, Interaction Analysis

> "investigates human activities, such as talk, nonverbal interaction, and the use of artifacts and technologies, identifying routine practices and problems and the resources for their solution. [13] "

The emphasis on how individuals interact with one another and with the resources at their disposal is of central interest to the current analyses. Furthermore, interaction analysis takes a situated perspective, in which a given behavior can only be understood in the context of the surrounding actions. As such, one way for framing the current analyses is towards determining an appropriate means for identifying a unit of analysis that can be used to develop meaningful segments of interactions. Additionally, Interaction Analysis is firmly rooted in human-based ethnographic video analysis. Accordingly the current effort aims to create automated approaches for conducting rich interaction analysis.

Given the apparent ties to Interaction Analysis, the connection to Embodied Cognition [16] should be apparent. However, invoking embodied cognition goes one step further by actively suggesting that there is a dependence on the body in order to fully engage in the process of cognition. The multimodal behavioral representations that we construct may serve as a window into better understanding the connections between body and mind. More specifically, the multimodal behavior representation may provide a glimpse into the behaviors that foster meaningful insights during a given task, and those that seem to detract from learning.

The current work is also informed by the theory of Epistemological Frames. [20,23] describe the connection between multimodal behavior and epistemology. Moreover, their work suggests that a student's body position, gaze, speech, and gesticulations can provide a lens into how a students is approaching a given task. It can also reflect the student's understanding of how they think they are expected to approach a given task. As an example of this, [23] identified four epistemological frames that could be used to model how students are approaching group work with their peers. Each of those epistemology frames could be reliably coded using multimodal data. Given that these multimodal markers can be a good indicator of student epistemology it would follow that identifying these moments in the context of learning could provide a useful way for denoting different "phases" of an activity, and subsequently, be useful for data segmentation.

Finally, this study is informed by work on Cognitive Disequilibrium which suggests that students experience learning when they enter into states of confusion [8]. Furthermore, the researchers have found that cognitive disequilibrium and feelings of frustration are mirrored in individual body movements as "fractal scaling" or pink noise [7]. If affective states represent such an important aspect of learning, and cognitive disequilibrium can be evidenced in body movement, it seems appropriate to explore the possibility of using facial expressions, a proxy for affective state, as a means for providing semantically relevant segments.

2.2 Multimodal Segmentation

There are certainly instances of prior research that utilize multimodal data to automatically segment data. Several examples exist within the television domain where researchers wish to utilize audio/visual data to automatically detect commercials, or naturally segment news broadcasts [5,10]. Other researchers have used multimodal data to conduct better segmentation of spoken text [6,19]. However, we know of no prior research that looks to use changes in facial expression and/or changes in body pose to automatically segment multimodal process data, especially when used in the context of trying to better understand and predict constructs related to student learning and cognition.

3. METHODS

To provide the reader with additional context, we briefly describe the research participants and the task that they completed before entering into a discussion of the analyses and results.

The task was motivated by prior work in Constructionism[17]. Students were given common household materials: one paper plate, 4 ft. of garden wire, four drinking straws and five wooden Popsicle sticks. The objective was to use the materials provided to create a structure that could support a weight of approximately half a pound. Participants were also asked to support the weight as high off the table as possible.

Our population of students consisted of twelve 9th- through 12th-grade students and eight undergraduate students. Pairs of students were randomly assigned to either use example- or principle-based reasoning, after controlling for prior education experience. Thus, each condition had six high school students and four undergraduate students. In the example-based condition, students generated three example structures from their home, community or school in order to motivate their design. In the principle-based condition, students identified three engineering principles that conferred strength and stability to a ladder, an igloo and a bridge, before embarking on the building task.

The data capture environment included: a Kinect sensor – for capturing audio, gesture and video; a high resolution web camera - to record how students moved the different materials; and an electro-dermal activation sensor – for measuring stress and/or arousal. All sensor data was synchronized through the data collection software, and also verified by a research assistant.

3.1 Activity Sequence

The overall flow of activities that students completed included: a pre-test; an intervention, i.e. one of the two conditions; a preliminary design drawing; a hands-on building activity; post-test; and reflection (Figure 1).

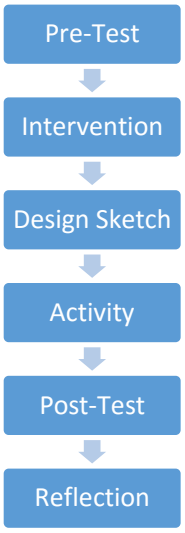

Figure 1. Overall study design

3.2 Learning-Related Constructs

As previously noted, our analysis was conducted against three different learning-related constructs: success, learning and experimental condition.

Success. Success is based on the performance of each pair's structure. Structures that are able to support the weight for more than thirty seconds, are coded as being successful. Structures that are unable to support the weight for more than thirty seconds are coded as being unsuccessful.

Learning. Learning is based on learning gains from pre-test to post-test. The pre- and post-tests asked students to identify important concepts in engineering design. Students who experienced positive learning gains were identified as having learned. Students who did not experience positive learning gains were said to have not learned.

Experimental Condition. As previously noted, students participated in one of two interventions. The first intervention, example-based reasoning, involved the students working together to come up with a design based on a real-world object from their home, community or school. The second intervention, principle-based reasoning, involved the students working together to come up with a design based on engineering principles. From prior work, we've observed that students in the principle-based reasoning condition are more successful and experience higher learning gains [26]. We've also observed that students use a better engineering design process when employing principle-based reasoning strategies. Hence, even though there is a correlation between experimental condition, success and learning, we are independently interested in identifying the multimodal behaviors associated with the two experimental conditions.

3.3 Data

The analyses included in this paper are drawn from the following multimodal data.

Audio data. Data was derived from a combination of audio channels from an overheard web camera, and from the Xbox Kinect sensor. Custom software was developed based on the Carnegie Mellon University (CMU) Sphinx Speech Recognition Toolkit [14]. Specifically, the source code was modified to leverage the program's voice activity detection feature. Voice activity detection is an automated means for determining when voice-based audio is being generated. Several speech recognition software solutions contain some variant of voice activity detection. The custom software provided voice detection start and stop times for all of the audio channels. Audio was considered to be present if either of the audio sources detected a voice, within a given second of time. Thus the final format of this data is a binary representation. Every second of the activity is labeled with a zero or one, for the absence or presence of audio at that time stamp. Because the audio channel captured sound from both participants this piece of data is the same for each person in a dyad.

Hand/wrist movement. Hand/wrist movement data was generated from the Xbox Kinect sensor. A custom built application was used to store three dimensional data for twelve upper body joints. The application uses native features available from the Kinect for Windows SDK, specifically, the ability to conduct skeletal tracking in the seated position. The custom application stores the data at 10 Hz. From the file generated, we utilize only the left and right wrist, hand and elbow data points. For each successive pair of data points we compute the angular displacement for the vectors that connect: left wrist and left hand; left wrist and left elbow, right wrist and right hand; right wrist and right elbow. The eventual angular displacement that is recorded is an average of the four angular displacements. Using angle as the means for comparison reduces biases introduced by participants having different sized bodies and limbs. Accordingly, for each tenth of a second in time we have stored the angular hand/wrist displacement. Finally, the cumulative angular displacement for each second of time is stored for each participant.

Electro-dermal Activation (EDA). Electro-dermal activation (also referred to as galvanic skin response and/or skin conductance) readings were captured at 8 Hz. Processing electro-dermal activation data involved controlling for individual differences in variance, as well as individual differences in stress response. In practice, this was achieved by collecting baseline data as students completed the task of counting down by 7.We will refer to this as the "math" stress test. As additional baseline data, students also completed a Stroop test, and had their electro-dermal activation recorded during non-task related activities. As before, each data point was time-stamped with the local date and time. Each data point was then transformed into an index value by subtracting the mean from the "math" stress test, and then dividing by the standard deviation of the "math" stress test data for that student. When we compared electro-dermal activation index values across the different activities, there were no statistically significant differences between experimental conditions for the baseline data, the Stroop test, or the math test. However, across the intervention, design phase and the building activity, differences were statistically significant. This provided validation that this normalization was effective. Electro-dermal activation index for each second of time was determined by taking the average electro-dermal activation index for a given second in time.

Facial Expression: Facial expression was extracted from frontal images using FACET SDK [15]. The facial expressions included: joy, anger, sadness, surprise, fear, contempt, disgust, confusion, and frustration. Evidence values for each of the facial expressions were used as an indicator for the presence and/or absence of each facial expression. Data points were recorded and stored at 1-second increments. All facial expressions were used for affect-based segmentation, while only confusion was used when quantifying multimodal behaviors. This will be described in more detail in the following sections.

Head Pose: Head pose estimation was determined from frontal images using a custom Constrained Local Neural Fields [2] application. The software provided pitch, roll and yaw at 1-second increments. Because participants have oppositely signed yaw values when they are looking at each other (they were seated side by side), the yaw values were transformed so that positive yaw corresponds to looking toward their partner and negative yaw corresponds to looking away from their partner. Finally, pitch and yaw are stored separately in the resultant vector.

3.4 Algorithm

The approach follows our previous work [25,27] on analyzing design strategies and success in hands-on engineering tasks. Here we significantly extend that work by incorporating different data segmentation strategies. A visualization of the general algorithm is provided in Figure 2 and Figure 3. A summary of each step is included below.

Time-stamp. The first step of extracting process data is to ensure that all data is properly time-stamped. This provides a means for synchronizing across the different modalities and results in a 7-dimensional matrix for each student. The dimensions of this matrix are time, confusion evidence, pitch, yaw (in terms of looking away from or towards one's partner), electro-dermal activation index, hand/wrist displacement and audio.

Segment. The time-stamped data is then segmented. We will refer to these segments as "data segments." Segmentation was adopted to help smooth the data, and provide units of analysis that are meaningful given the learning constructs of interest.

For this study we compare four different segmentation strategies: human-based, pose-based, affect-based and fixed-window.

Human-based segmentation creates a new segment every time a pair's structure is tested. This approach is based on interpreting testing as an instance in which at least one person in the pair is eliciting feedback that will update the students on the current stability of their structure. Testing usually takes the form of a team member placing the weight on the structure. Under this segmentation strategy dyads had the same number of "data segments." Finally, this approach is informed by traditional Interaction Analysis techniques [13] and our prior work [24,25,27].

Pose-based segmentation creates a new segment every time a given student changes the direction of their head pose. In more precise terms, a new segment was defined as a change in the sign of the pitch and/or yaw. A minimum pose duration of 8 seconds is imposed, though this limit is seldom utilized. Recall that this approach is motivated, in part, by Epistemological Framing [20,23].

Affect-based segmentation creates a new segment every time there is a change in the most evident facial expression. Once again, a minimum segment duration of 8 seconds is imposed. A primary motivation for this segmentation strategy is research on the importance and manifestations of cognitive disequilibrium [7,9]

Fixed-window segmentation creates a new segment at every second of time. In conducting our analysis, we tested varying fixed-window sizes (e.g. 4, 8, 10 and 30 seconds) but found that 1 second windows worked the best. Because the data is preprocessed to be grouped on a per second basis, fixed-window segmentation requires no additional processing at this step.

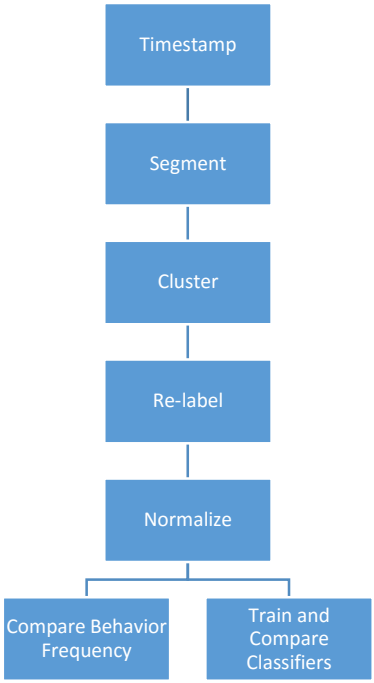

Figure 2. General algorithm

As a whole, the segmentation process serves to smooth the data. Instead of having to take into account each of the spikes and troughs that may emerge from any of the data streams, segmentation allows us to look at broader trends. Noise reduction is also achieved during the forthcoming clustering step.

Segmentation always resulted in a single value for each data stream. For the audio data the value is the speech fraction. For the hand/wrist movement data, the value is the average total angular displacement during that "data segment." For electro-dermal activation, the value is the average index value during that particular "data segment." Similarly, for head pose estimation, and confusion evidence, the values indicate the average pitch, yaw, and confusion evidence.

Cluster. After the segmentation step, there are hundreds to thousands of unique "data segments." Some of these will be very similar to one another, only differing by an infinitesimal amount, while others will vary quite extensively from one another. The goal of clustering is to identify natural groupings among the various "data segments" and ultimately provide a common set of states, or behaviors, by which to compare students. However, before proceeding with clustering, we first do data standardization. Namely, we adjust each value, such that all of the data in a given column has a mean of zero and a standard deviation of one. This process eliminates bias in clustering, by ensuring that each column contributes equally to the distance metric, which in the case was Euclidean distance. After standardizing the data, we used K-Means clustering, with a Euclidean distance metric, to group the data points into a set of four clusters that place each "data segment" with the other "data segments" that it is most similar to. Four clusters was used based on our prior experience with this dataset. Specifically, in [27] we found through human observation that four clusters provided the most meaningful semantic meaning.

Several iterations were completed to avoid local maxima. Once each "data segment" has been grouped with similar "data segments," each cluster, or group, can be described based on the cluster centroid value. These values provide the basis for determining common behavioral practices in later sections.

Note: We tried variable sized clusters, allowing for each segmentation strategy to use the number of clusters that conferred the greatest predictive accuracy. However, there was very little difference in the results. Hence, we used the same number of clusters for each segmentation strategy, as to reduce algorithmic variation, and simplify the analysis.

Re-label. All "data segments" that are put into the same cluster are given the same name and value. Accordingly, each student's sequence of "data segments" is represented as a list of clusters.

Normalize. In the normalization step, each student's data is reduced to a four dimensional vector that features the frequency that cluster was used by that student (L-1 normalization).

Compare Behavior Frequency. After L-1 normalization, the next step is to compare behavior frequency data across the three metrics of interest: success, experimental condition; and learning. The comparisons are based on two-tailed t-tests along each of the individual clusters for a given segmentation strategy.

Train and Compare Classifiers. The frequency values are used to train classifiers for predicting the three metrics of interest: success, experimental condition; and learning. In this study we trained a support vector machine, with a linear kernel. Sixteen fold leave-one out training and testing was completed to determine the effectiveness of each approach. Comparisons were based on F-score from the test data, as compared to a majority class classifier.

4. RESULTS

4.1 Correlations between cluster frequency and learning constructs

We begin our presentation of the results with a discussion of the cluster frequencies, paying particular attention to how well each cluster correlated with success, learning and experimental condition (Tables 1 – 4). We report Pearson correlation values, and p-values based on two-tailed t-tests, with 15 degrees of freedom. Specifically, for each construct we compute the level of correlation and the probability that successful students used a given cluster (multimodal behavior) with greater frequency than unsuccessful students, for example. In each table "*", "**" and "***" correspond to $p < 0.05$, 0.01 ad 0.001, respectively.

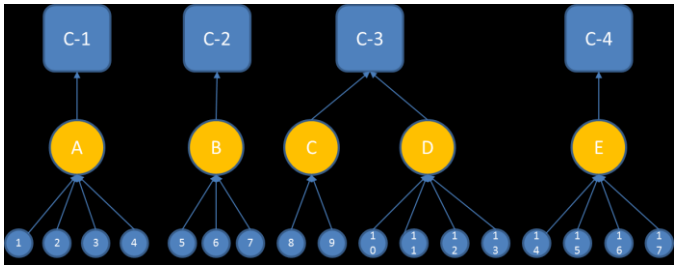

Figure 3. Pictorial representation of the data algorithm. Numbers 1 through 17 represent different seconds in time. A through E represent "data segments" that are defined based on the individual timestamps associated with them. C-1, C-2, C-3, C-4 represent multimodal behavior clusters. The clusters are based on using K-means to group "data segments".

Table 1. Pearson r values and p-values for cluster usage by construct for fixed window segmentation

Fixed Window			
Cluster	Success	Condition	Learning
1	0.50*	0.44*	0.62*
2	0.48*	0.44	0.29
3	0.22	0.11	0.25
4	0.34	0.11*	0.46

Table 2. Pearson r values and p-values for cluster usage by construct for affect-based segmentation

Affect			
Cluster	Success	Condition	Learning
1	0.50*	0.46*	0.67*
2	0.49	0.45	0.41
3	0.35	0.32*	0.44
4	0.25	0.12	0.25

Table 3. Pearson r values and p-values for cluster usage by construct for pose-based segmentation

Pose			
Cluster	Success	Condition	Learning
1	0.44*	0.55*	0.90***
2	0.38	0.48*	0.56*
3	0.19	0.25	0.25
4	0.19	0.25	0.25

Table 4. Pearson r values and p-values for cluster usage by construct for human-based segmentation

Human			
Cluster	Success	Condition	Learning
1	0.42*	0.37	0.35
2	0.47	0.25	0.25
3	0.52	0.29	0.22
4	0.48	0.32	0.25

Substantively, we see that cluster 1 from affect-, pose- and fixed-window based segmentation, significantly correlate with success, learning and experimental condition (see Table 1, Table 2 and Table 3) all with relatively large effect sizes (between 0.44 and 0.90). Additionally, cluster 2 from fixed-window segmentation (Table 1), cluster 3 from affect-based segmentation (Table 2) segmentation, and cluster 1 from human-based segmentation all correlate with success (Table 4). We also see that cluster 2 from pose-based segmentation correlates with experimental condition and learning (Table 3).

Looking at the Pearson r values we see that across learning constructs, affect- and/or pose-based segmentation are typically of equal or greater correlation than fixed-window and human-based

segmentation. This is especially true for experimental condition and learning.

In summary, then, affect-, pose- and fixed-window-based segmentation result in cluster centroids that offer greater correlation with the three learning constructs of interest, than human-based segmentation. Furthermore, we see the pose-based segmentation provided the greatest number of significant correlations among the cluster centroids and learning constructs, all with relatively large effect sizes.

4.2 Cluster Interpretation

Now that we have seen that the clusters have significance for drawing correlations between multimodal behaviors and the three learning constructs, we now turn to briefly analyzing the behaviors associated with the clusters.

To simplify the interpretation of each cluster centroid, and facilitate comparison, we transformed the cluster centroid values into relative values along each dimension, as follows:

High: cluster centroid value more than a half of the standard deviation above the mean.

Medium: cluster centroid value more than a tenth of a standard deviation above the mean, but less than half of a standard deviation above the mean.

Low: cluster centroid value less than a tenth of a standard deviation above the mean (includes values that are less than the mean).

Head pose estimation was transformed into a single value that describes overall gaze. A prefix of "u" denotes looking up, while a prefix of "d" denotes down. "p" refers to looking toward their partner, while "a" denotes looking away from their partner.

Table 5 describes cluster 1 for each segmentation technique. Recall that cluster 1 usage frequency generally correlated with success, learning and experimental condition.

Table 5. Characteristics for Cluster 1 from each approach. In the Pose dimension, "d" denotes down, "u" denotes up, "p" denotes towards their partner, "a" denotes away from their partner

	Confusion	*Pose*	*EDA*	*Gesture*	*Audio*
Fixed Window	medium	d p	low	medium	high
Affect	medium	d p	low	medium	high
Pose	medium	u p	high	medium	medium
Human	medium	u p	low	medium	high

From Table 5, we see that cluster 1 from fixed window segmentation and affect-based segmentation constitute a very similar multimodal behavior that is typified by actively engaging the task while looking in the direction of one's partner. Engagement is observed by the high audio, and medium gesticulation values. In fact, all of the cluster 1's share above average hand/wrist movement, and above average audio, suggesting that the participants are engaged. This is in contrast to a cluster that may be typified by low gesticulation and low audio. Deviating away from a behavior that is characterized by active engagement, appears to be associated with less successful designs, lower learning gains, and example-based reasoning. This result mirrors what was observed in our prior work [27].

To summarize, then, our comparison across the four different approaches, in terms of correlations between multimodal behaviors to the three learning-related constructs, shows that affect- and pose-based segmentation appear to be equally as effective as fixed-window segmentation, and superior to human-based segmentation. Part of this correlation appears to be couched in detecting "engaged" behavior. Additionally, pose-based segmentation stands out as a particularly good strategy for examining these correlations. In the next section we examine if this is also true for making predictions about success, learning and experimental condition.

4.3 Predictive power of segmentation approach

An SVM classifier with a linear kernel was trained using the cluster centroid frequencies for each student, and each approach. Precision and recall were computed for each class, against ground truth. Those values were then used to compute the average F-score as a point of comparison across the four different approaches. Results are summarized in Table 6.

Table 6. Average F-score by learning construct and segmentation approach (baseline denotes a majority class single assignment classifier). Bold denotes high performer.

Approach	*Condition*	*Learning*	*Success*
Pose	**0.73**	0.26	**0.62**
Fixed Window	0.68	0.54	0.56
Affect	0.68	0.11	0.55
Human	0.52	**0.64**	**0.62**
Baseline	0.33	0.45	0.38

From Table 6, we see that for predicting each student's experimental condition, pose-based segmentation results in the highest average F-score, clearly outperforming the majority class classify, human-based segmentation and fixed-window segmentation. Additionally, pose-based segmentation matched human-based segmentation for predicting success. However, as it relates to "learning," human-based segmentation and fixed-window segmentation significantly outperformed affect- and pose-based segmentation. Furthermore, affect- and pose-based segmentation failed to outperform the majority class classifier for "learning."

5. DISCUSSION

The current analyses were motivated by a desire to identify data segmentation strategies that could provide comparable results to those achieved using human-based segmentation and fixed-window segmentation. More specifically, we hypothesized that affect- and pose-based segmentation would be as effective, or more effective, than human-based and fixed-window segmentation, for correlating and predicting the learning-related constructs: success, learning and experimental condition.

The correlation analysis made salient the relative effectiveness of affect- and pose-based segmentation for drawing connections between multimodal behaviors and all three of the learning constructs. In particular, the analysis revealed that fixed-window, affect-based, and pose-based segmentation, all produced at least one cluster that constituted a multimodal state of "active engagement." Students who performed better, learned more and came from the principle-based reasoning condition were more likely to evidence this multimodal state of "active engagement." In this way, the affect- and pose-based segmentation strategies proved to be just as effective as the fixed-window segmentation strategy. Additionally, pose-based segmentation outperformed fixed-

window segmentation in that it had a second cluster centroid that correlated with experimental condition and learning. As such, the hypothesis that affect- and pose-based segmentation can be equally as effective for correlating among multimodal behaviors and learning-related constructs appears to be confirmed.

On the topic of predicting each of the learning-related constructs, the results were slightly different. The classifier built using the pose-based segmentation approach was the best for predicting both success and experimental condition. And while affect-based segmentation was generally the worst for all three of the learning constructs, it did outperform both human-based annotation and the majority class classifier for predicting experimental condition. However, whereas pose- and affect- based segmentation demonstrate great promise for predicting success and experimental condition, they were both extremely ineffective for predicting learning. As such, it may be that learning, as a construct, is harder to predict than success and experimental condition. Put differently, the link between multimodal behaviors and learning, may not be as easily represented, because learning is a process that often involves cognitive processes that do not map as easily onto multimodal behaviors.

As we consider the reasons for why affect- and pose-based segmentation were effective in this study, we first return to the idea of variable length "data segments." One of our conjectures for the utility of human-based segmentation is that it results in variable length segments, as opposed to fixed length segments. Having variable length segments seems to more closely mirror our human experiences, as a lot of how humans learn and operate, is situational [4]. Accordingly, it may be that affect- and pose-based segmentation are benefiting from improvements related to variable length segments.

Another area for consideration, is the level of granularity across the four approaches (Table 7). When we examine the total number of segments for each approach, we see that human-based segmentation has the fewest, affect-based has the second fewest, pose-based has the second most, and fixed-window segmentation has the most. This puts affect- and pose- based segmentation near the middle of the spectrum and may help in providing enough granularity that correlations can be accurately observed. At the same time, being at the middle of the spectrum may also provide enough aggregation that there is somewhat of a semantic meaning to segments, beyond an aggregation of otherwise disconnected moment-by-moment actions. In our ongoing research we intend to investigate this further.

Table 7. Number of segments by segmentation approach

Approach	*Number of Segments*
Human	236
Affect	768
Pose	5995
Fixed Window	24569

Finally, prior work on levels of abstraction may offer some insights here as different constructs are evidenced over markedly different time scales [1] and at different levels of abstraction [21]. [1] describes learning as happening on the order of days and months, whereas something like success, or epistemology can more easily be inferred through more fine-grained data. [21] makes a similar point that moving across different levels of abstraction can be quite challenging, may require larger data sets and novel approaches for dealing with uncertainty.

6. FUTURE WORK

In future work we plan to examine the efficacy of affect- and pose-based segmentation on a larger data set. As part of this study, we will be curious to look a more fine grain differences in pose and affect that we could not reliably examine with the current data set, due to its relatively small sample size. For example, the current affect-based segments were based on changes in facial expression evidence. With a larger dataset a may be possible to pay more attention to actions units, and/or look at a larger array of possible poses. With a larger dataset, we would also explore modeling how certain multimodal behaviors may actually be useful for predicting the subsequent pose, and vice versa. Additionally, because pose-based segmentation appeared to be particularly promising, future analyses will study dyadic pose-based interactions, and more in-depth differentiation of poses.

7. CONCLUSION

Our primary objective for this paper was to explore the possibility of using affect- and posed-based segmentation as an alternative to human-based segmentation and fixed-window segmentation. We correctly hypothesized that affect- and pose-based segmentation would be on par with the two more traditional techniques for studying correlations between multimodal behaviors and three constructs related to a given learning experience: success, learning and experimental condition. Pose- and affect-based segmentation both performed extremely well for defining correlations between multimodal behavior and the three constructs. Moreover, pose-based segmentation stood out as extremely relevant for this study. Similarly, on the prediction task, pose-based segmentation, proved to be a top performer for predicting success and experimental condition, easily outperforming the majority class baseline and the two traditional forms of data segmentation. However, affect- and pose-based segmentation performed quite poorly for predicting learning, significantly lagging behind human-based segmentation, which proved to be the best for this particular construct. Nonetheless, along nearly all of the dimensions of comparison affect- and pose-based segmentation performed quite well, suggesting that they may present a viable means for automatically segmenting process data in a way that results in variable length segments, and that maintains semantic meaning. Finally, while this process was presented in the context of student learning, it may have useful implications in other domains for which variable length segments are common.

8. ACKNOWLEDGEMENTS

A portion of this research was made possible through the financial support of the University of Southern California's Provost's Signature Fellowships and the Stanford University Vice Provost of Graduate Education.

9. REFERENCES

1. John Anderson. 2002. Spanning seven orders of magnitude: a challenge for cognitive modeling. *Cognitive Science* 26, 1, 85–112. http://doi.org/10.1207/s15516709cog2601_3
2. Tadas Baltrušaitis, Peter Robinson, and Louis Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. *Proceedings of the IEEE International Conference on Computer Vision*, 354–361. http://doi.org/10.1109/ICCVW.2013.54
3. Paulo Blikstein and Marcelo Worsley. in press. Multimodal Learning Analytics: a methodological

framework for research in constructivist learning. *Journal of Learning Analytics*.

4. JD Bransford, AL Brown, and RR Cocking. 2000. How people learn. Retrieved July 15, 2014 from http://www.csun.edu/~SB4310/How People Learn.pdf
5. Mattia Broilo, Eric Zavesky, Andrea Basso, and Francesco G B De Natale. 2010. Unsupervised Event Segmentation of News Content with Multimodal Cues. *Proceedings of the 3rd International Workshop on Automated Information Extraction in Media Production*, ACM, 39–44. http://doi.org/10.1145/1877850.1877862
6. Lei Chen. 2006. Incorporating Gesture and Gaze into Multimodal Models of Human-to-human Communication. *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Doctoral Consortium*, Association for Computational Linguistics, 211–214. http://doi.org/10.3115/1225797.1225798
7. S D'Mello, Rick Dale, and Art Graesser. 2012. Disequilibrium in the mind, disharmony in the body. *Cognition & emotion*, 1–28. Retrieved May 06, 2014 from http://www.tandfonline.com/doi/abs/10.1080/02699931.2011.575767
8. Sidney D'Mello and Art Graesser. 2012. Dynamics of affective states during complex learning. *Learning and Instruction* 22, 2, 145–157. http://doi.org/10.1016/j.learninstruc.2011.10.001
9. SK D'Mello, B Lehman, and Natalie Person. 2010. Monitoring affect states during effortful problem solving activities. *... Journal of Artificial Intelligence in ...*. Retrieved May 19, 2014 from http://iospress.metapress.com/index/V9200K51R3820T58.pdf
10. Ling-Yu Duan, Jinqiao Wang, Yantao Zheng, Jesse S Jin, Hanqing Lu, and Changsheng Xu. 2006. Segmentation, Categorization, and Identification of Commercial Clips from TV Streams Using Multimodal Analysis. *Proceedings of the 14th Annual ACM International Conference on Multimedia*, ACM, 201–210. http://doi.org/10.1145/1180639.1180697
11. Joseph F Grafsgaard. 2014. Multimodal Analysis and Modeling of Nonverbal Behaviors During Tutoring. *Proceedings of the 16th International Conference on Multimodal Interaction*, ACM, 404–408. http://doi.org/10.1145/2663204.2667611
12. Larry Johnson, Samantha Adams, Malcolm Cummins, Victoria Estrada, Alex Freeman, and Holly Ludgate. 2013. The NMC horizon report: 2013 higher education edition.
13. Brigitte Jordan and Austin Henderson. 1995. Interaction Analysis: Foundations and Practice. *The Journal of the Learning Sciences* 4, 1, pp. 39–103. Retrieved from http://www.jstor.org/stable/1466849
14. Kai Fu Lee, Hsiao Wuen Hon, and Raj Reddy. 1990. Overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 38, 35–45. http://doi.org/10.1109/29.45616
15. Gwen Littlewort, Jacob Whitehill, Tingfan Wu, et al. 2011. The computer expression recognition toolbox (CERT). *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, 298–305. http://doi.org/10.1109/FG.2011.5771414
16. RE Núñez, LD Edwards, and JF Matos. 1999. Embodied cognition as grounding for situatedness and context in mathematics education. *Educational Studies in Mathematics*, 39, 45–65. Retrieved March 17, 2014 from http://link.springer.com/article/10.1023/A:1003759711966
17. Seymour Papert. 1980. *Mindstorms: Children, computers, and powerful ideas*. Basic Books, Inc.
18. Roy Pea. 2014. THE LEARNING ANALYTICS Workgroup A Report on Building the Field of Learning Analytics.
19. Francis Quek, David McNeill, Robert Bryll, et al. 2002. Multimodal Human Discourse: Gesture and Speech. *ACM Trans. Comput.-Hum. Interact.* 9, 3, 171–193. http://doi.org/10.1145/568513.568514
20. Rosemary S. Russ, Victor R. Lee, and Bruce L. Sherin. 2012. Framing in cognitive clinical interviews about intuitive science knowledge: Dynamic student understandings of the discourse interaction. *Science Education* 96, 4, 573–599. http://doi.org/10.1002/sce.21014
21. Stefan Scherer, Michael Glodek, Georg Layher, et al. 2012. A generic framework for the inference of user states in human computer interaction. *Journal on Multimodal User Interfaces* 6, 3-4, 117–141.
22. Stefan Scherer, Marcelo Worsley, and Louis-Philippe Morency. 2012. 1st international workshop on multimodal learning analytics. *ICMI*, 609–610.
23. Rachel E. Scherr and David Hammer. 2009. Student Behavior and Epistemological Framing: Examples from Collaborative Active-Learning Activities in Physics. *Cognition and Instruction* 27, 2, 147–174. http://doi.org/10.1080/07370000902797379
24. Marcelo Worsley and Paulo Blikstein. 2013. Towards the Development of Multimodal Action Based Assessment. *Proceedings of the Third International Conference on Learning Analytics and Knowledge*, ACM, 94–101. http://doi.org/10.1145/2460296.2460315
25. Marcelo Worsley and Paulo Blikstein. 2014. Analyzing Engineering Design through the Lens of Computation. *Journal of Learning Analytics* 1, 2, 151–186.
26. Marcelo Worsley and Paulo Blikstein. 2014. Assessing the Makers: The Impact of Principle-Based Reasoning on Hands-on, Project-Based Learning. *Proceedings of the 2014 International Conference of the Learning Sciences (ICLS)* 3, 1147–1151.
27. Marcelo Worsley and Paulo Blikstein. 2015. Leveraging Multimodal Learning Analytics to Differentiate Student Learning Strategies. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, ACM, 360–367. http://doi.org/10.1145/2723576.2723624
28. Marcelo Worsley. 2012. Multimodal learning analytics: enabling the future of learning through multimodal data analysis and interfaces. *Proceedings of the 14th ACM international conference on Multimodal interaction*, 353–356.