Automatic Detection of Nonverbal Behavior Predicts Learning in Dyadic Interactions

Andrea Stevenson Won, Jeremy N. Bailenson, and Joris H. Janssen

Abstract—Nonverbal behavior can reveal the psychological states of those engaged in interpersonal interaction. Previous research has highlighted the relationship between gesture and learning during instruction. In the current study we applied readily available computer vision hardware and machine learning algorithms to the gestures of teacher/student dyads ($N = 106$) during a learning session to automatically distinguish between high and low success learning interactions, operationalized by recall for information presented during that learning session. Models predicted learning performance of the dyad with accuracies as high as 85.7 percent when tested on dyads not included in the training set. In addition, correlations were found between summed measures of body movement and learning score. We discuss theoretical and applied implications for learning.

Index Terms—Natural data set, machine learning, gesture recognition, collaborative learning

1 Introduction

Attending to nonverbal behavior is a key component of teaching and learning. Beyond the meaningful gestures that explicitly support content [1], body movements relate to the attitudes of the participants and outcomes of interactions. Gesture and posture in educational contexts have thus been examined for what they may reveal about teaching and learning (for a review, see Roth, [2]). In the following pages we review previous work investigating the role of nonverbal behavior in teaching and learning. We discuss related work on automatically detecting affect and other mental states. We then describe a new study utilizing computer vision and machine learning to predict the outcome of teaching/learning interactions, based on the general tracked body movements of the interactants.

1.1 Gestures in Teaching and Learning

A number of studies have examined the relationship between students' nonverbal behavior and attentiveness and comprehension. For example, students' nonverbal behaviors have been recorded and correlated with observers' reports to predict students' levels of engagement, with the goal of developing automated systems that could help predict and assist learning. Mota and Picard [3] used a pressure sensitive chair to track the posture cues of children performing a learning task at a desktop computer, relating these cues to observers' ratings of the children's levels of interest. Static postures and sequences of postures were tracked with the goal of developing automatic detection systems that could be used both to refine current understanding of behavior during learning, and to allow for the development of learning tools. In 2008 Dragon et al. [4] observed students using a computer tutor, and coded physical and affective behaviors. A separate group of researchers then used this data to design an intelligent tutoring system that used posture and facial feature tracking to detect learner affect and adjust the computerized tutor accordingly to optimize learning [5], [6]. These efforts provide a foundation for further research to improve learning via detecting nonverbal behavior.

Since successful communication between teacher and student is one critical component of the learning process, development of teacher/student rapport via *synchronous* nonverbal behavior has also been examined in a teaching context. In a 1976 study using human coders, LaFrance and Broadbent [7] recorded classroom behavior, noting whether students in small classroom settings a) mirrored (copied their teacher's movements on the other side of their body; for example, raised the right hand when the teacher raised his or her left hand) b) matched (copied their teacher's gestures using the same side of their body; for example raised the right hand when the teacher raised his or her right hand) or c) had incongruent behavior (not perceived by observers to echo that of the teacher in any way). The researchers found a correlation between synchrony (either mirrored or matched movements between teacher and student gestures) and students' self reports of involvement and rapport. Similarly, Bernieri [8] had coders rate perceived movement synchrony (described as "simultaneous movement, tempo similarity, and smoothness") of high school students in teaching/learning dyads. The synchrony of the teaching interaction correlated with students' self-reported rapport. In a recent sudy [9], reciprocal gestures (coded by humans) between teachers and students engaged in a language task not only correlated with reported rapport, but also with higher student quiz scores.

Traditionally, research on nonverbal behavior has taken advantage of humans' top-down observational ability to

- *A.S. Won and J.N. Bailenson are with the Department of Communication, 450 Serra Mall, Stanford University, Stanford, CA 94305-2050. E-mail: {aswon, bailenson}@stanford.edu.*
- *J.H. Janssen is with Sense Observation Systems, Lloydstraat 5, 3024EA Rotterdam, The Netherlands. E-mail: joris@sense-os.nl.*

Manuscript received 8 Feb. 2013; revised 3 May 2014; accepted 21 May 2014. Date of publication 8 June 2014; date of current version 23 July 2014. Recommended for acceptance by R.A. Calvo. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2014.2329304

perceive gestalt phenomena such as synchrony by using human observers to code recorded data, as seen in many of the studies described above. However, hand-coding nonverbal behavior is extremely labor intensive. Data must be recorded and coded post-task, or observers must watch the participants in real time. In addition, observers bring their own biases to the interpretation of behaviors, and may be influenced by other factors such as facial expressions or the content of the conversation. Thus, this kind of monitoring is expensive and slow, and it is difficult to process large amounts of data quickly or to evaluate various channels independently. In the current study, leveraging the automatic detection and analysis of gesture allowed us to examine large data sets. The research described in this paper seeks to build on previous work by incorporating a more bottom-up method of assessing the importance of body movements in a naturalistic environment.

1.2 Automatic Detection and Analysis of Gesture

As an alternative to human observation and coding, research on automatically detecting and analyzing nonverbal behavior to predict emotions and other affective states has proceeded on many fronts over the past few decades.

Combining facial expression with other modalities, Meservy et al. [10] used head and hand movements to detect deception using video recordings. Similarly, Karpouzis et al. [11] used multimodal signals, including facial expression, hand gestures, and prosody (pitch and rhythm of voice) to detect naturally occurring emotion during an interaction between a human and an embodied agent. Research using facial tracking alone to predict outcomes includes detecting and identifying facial expressions [12]; distinguishing between similar facial expressions, such as frustrated or delighted smiles [13]; and identifying the tendency of participants to make mistakes in a task [14]. Finally, recent work combines the affective measurements from the individuals in an interaction to assess outcome [15].

Gestural and postural information, especially large-scale body movements, may be easier to access in natural conditions than other information. For example, facial expressions can be obscured by makeup, eyeglasses, or facial hair. Lighting conditions, occlusion, or head position may also make these expressions difficult to read. Ambient sound may confuse audio cues, and physiological signals may impose prohibitive constraints. Especially in nonlaboratory environments, large-scale movements can be an important addition to other modalities.

As Kleinsmith and Bianchi-Berthouze [16] and others [17] demonstrate in their assimilation of the literature, much of the current research on detecting affect has focused on facial expressions, speech, and physiological signals. However, assessing gross signals of body movement, such as gesture and posture, also hold great promise for predicting and interpreting affective states. Movements potentially indicate general states of mind that are not necessarily specific to the verbal content of the conversation, and indicate the continuous evolution of these states in dyads. Particularly in interpersonal interactions, gestures may provide complementary information to that derived from facial expressions [18] or tone of voice because body movements are not subject to the same degree of conscious control [19]. Gestures and body movements can alter how people conceptualize abstract concepts [20] and even their sense of their own dominance [21].

Body gestures specifically have been used to predict affect. Kapur et al. [22] used a six-camera system to capture the X, Y, Z positions of 14 markers on five participants enacting four emotions. Machine learning was then used to classify the recorded gestures as indicating sadness, joy, anger or fear. Another experiment used video analysis of enacted motions [23] to distinguish between joy, anger, pleasure and sadness. In this experiment, participants were asked to use repeated arm movements for each emotion, so the classification was based on expressivity, rather than on gestures particular to a specific emotion. As part of an experiment assessing interpersonal touch via haptic devices, Bailenson et al. [24] demonstrated that a simple, two-degree of freedom device could transmit emotions via hand movements. In addition to these enacted emotions, natural emotions have been detected using touch on a screen during game play, which was used as the input to allow the automatic discrimination of four emotional states [25]. Similarly, the movements of players in a video game that tracked body movements were captured and used to predict affective states, with accuracies comparable to those of human observers' predictions [26]. In another example using recorded videos of speakers, the speakers' automatically tracked gestures were used to predict online ratings of that video [27].

Nonverbal communication predicts a variety of outcomes in interpersonal interactions, sometimes using very short time periods (also called "thin slices") of interaction [28]. Some examples of thin slice prediction based on nonverbal behavior include the rate at which doctors were sued for malpractice based on the doctor's tone of voice during routine office visits [29]; ratings of teachers' bias based on ratings of their nonverbal behavior (but not their verbal behavior) when speaking to students [30]; and the overall and session level success of psychotherapy based on the automatic coding of patient/therapist gestural synchrony [31]. These experiments imply great potential for detecting affect, predicting outcomes, and providing feedback to alter the course of an interaction using very short glimpses of an interaction.

In the following study, we leverage current technologies to measure gesture to predict the outcome of a teaching-learning interaction.

1.2.1 Collecting and Processing Data

A key component of these kinds of automated systems is using computer vision to collect data. While collecting gesture automatically via computer is challenging, past research indicates that body movements, even when recorded as relatively coarse measures of movement, are indeed useful for assessing behavior. The history of point-light displays illustrates the amount of information available from very sparse amounts of nonverbal information. Since Johansson [32] first showed that human observers were capable of distinguishing biological motion using films of confederates wearing black clothing and light

markers on major joints, other studies have demonstrated that humans are able to identify gender [33], sexual orientation [34], depression [35], and emotion [36] from this kind of minimal information.

Computer vision systems that detect human motion have often relied on analogous systems of optical markers worn on the body to track movement over time. Similarly, new video game interfaces such as the Nintendo Wii or Sony Move require wands or other sensors in order for the users' motions to be captured. The disadvantage of these kinds of interfaces is that they require the user to wear or use particular devices, which can be intrusive or even distracting [37]. The recent trend of using active computer vision, for example, the infrared systems that are employed in the Microsoft Kinect, provides a compromise between accuracy and unobtrusiveness, thus increasing the range of possible applications. This can be seen in recent papers successfully using Kinect to detect specific gestures, such as pill taking, in naturalistic environments [38], [39].

1.2.2 Study Design

Previous work that has utilized automatic systems to examine nonverbal behavior in the domain of teaching and learning has often focused more on design than evaluation. Metrics often concentrated on particular aspects of either the teacher or the student's contributions [3], [4], [6], and few studies have combined an objective measure of learning with input from both. We describe an experiment determining whether the outcome of a teaching/learning interaction could be predicted using naturalistic body movements captured by commercially available video game hardware.

We conducted a study assessing the interaction between teachers and students in a naturalistic environment attempting to predict the outcome of a teaching/learning task. By using unobtrusive interfaces, we hoped to record naturalistic gestural and postural data. Using a large number of teacher/student pairs, we aimed to capture general information about teacher/learner interactions. Following previous research in education [40], we broke our data set into subsets of increasingly extreme high- and low-success pairs. This allowed us to identify behaviors that could be more apparent in extreme cases. In order to move beyond self-reported or observer-coded rapport or engagement, we administered a written memory test as a first step towards measuring learning outcomes. Finally, we limited our focus to body movements grouped by five body regions but avoided defining any specific gestures, for example, nodding the head to indicate agreement. For a description of gesture categories see Roth, [2]. In this way, we were able to examine regions of the body in an anatomically meaningful way (by grouping movements by the arms, legs, and torso/head regions) while avoiding analyses based on specific gestures.

2 Methods and Materials

2.1 Participant Population

An initial convenience sample of 160 participants (80 teacher-student pairs) was composed of undergraduate or master's students from a medium-sized West Coast university, ranging in age from 18 to 22. The sample was evenly divided between male and female participants, who were randomly assigned to mixed or same-sex pairs. Twenty-seven participant pairs were removed due to equipment failure, or lack of sufficient amounts of tracking data that was matched for both participants, leaving 53 pairs (52 women and 54 men across all pairs). Participants received either course credit or a 15-dollar gift card for their participation. All participants signed an informed-consent form before beginning any part of the experiment.

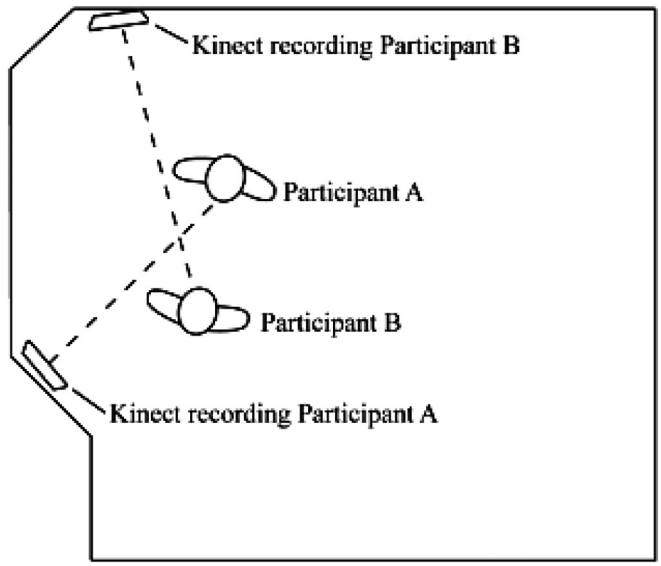

Fig. 1. A bird's eye view of the positions of the participants in relationship to the Kinect cameras mounted on the walls.

2.2 Apparatus

Two Microsoft Kinect [41] devices were used to capture participants' gestures and postures. These interfaces use an emitter and an infrared camera to capture body movement without requiring users to wear markers or hold any device. The cameras are integrated into a small panel, approximately $12 \times 6 \times 5$ inches in size, which weighs approximately 3 pounds. This small size and light weight allows the device to be wall-mounted or set on a tabletop.

Although the Kinect used in this experiment did not capture facial expressions or detailed hand movements, its tracking is noninvasive and can operate in low light conditions at distances between 1.22 to 3.65 meters. For this experiment, two Microsoft Kinect devices were used simultaneously.

2.3 Procedure

The first participant, or teacher, was directed to stand on a tape marker in the main lab room, facing the researcher. Kinect cameras were attached to the walls in front and to the right of each participant, so that each participant's movements were recorded without being obscured by his or her conversational partner, as shown in Fig. 1. This position was piloted to make sure the Kinects accurately tracked participants.

The researcher informed the teacher participant that he or she would be verbally taught a list of fifteen environmental principles, along with examples that helped to illustrate those principles (see Appendix A, which can be found

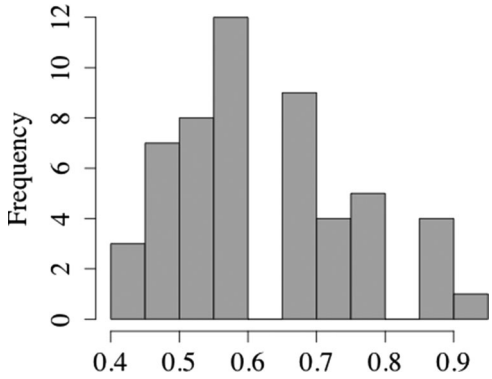

Fig. 2. Histogram of raw teacher free recall test scores, which is approximately normally distributed.

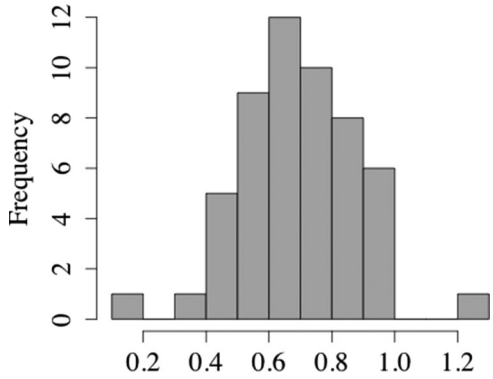

Fig. 3. Histogram of student free recall test scores transformed as percentage of teachers' scores. Note there are scores higher than 1.0 as some teachers may have mentioned items to the students but not listed those items on the recall test.

on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2014.2329304). After this brief teaching session, he or she would then teach that material verbally to a second participant.

The material to be taught was intended to be novel to most participants. This allowed us to examine a large number of teachers who would all be on approximately equal footing with each other, and also allowed us to examine body movements that were not specific to the material. If the material required the use of specific gestures (such as describing volume or direction) or if an experienced teacher had developed a routine for teaching the material, it might be more difficult to generalize from those movements.

After the researcher recited the fifteen principles and examples to the teacher participant, a second participant, the student, was brought in. The teacher and student were introduced. The experimenter then stated that the teacher would have five minutes to teach the list to the student, after which they would both take a brief written test. The experimenter then left the room and the participants began the interaction. Both participants were recorded by the Kinects throughout.

At the end of five minutes, or when signaled with a raised hand by the teacher participant, the researcher reentered the room and seated participants in separate rooms to take a written free-recall test. Participants listed as many of the principles that they had just been taught as they could remember. A free recall test was selected in order to maximize the variance in responses.

2.4 Measures

The 15 answers on the test were graded by two raters trained in the use of a key provided by the researcher. Since these were free responses, there was some subjectivity in the gradings. The rater's initial scores correlated at 0.89. To provide the most stable data set, two raters discussed the scores to resolve differences in order to come to a complete consensus on the entire data set. Student and teacher accuracy scores correlated at 0.57. In order to account for the teachers' ability to recount the principles, the students' scores were transformed as a percentage of the teachers' scores for each test. Thus, if the teacher scored eight on the free recall test, and their student also scored eight, the student would receive a score of 1.0 (i.e., 100 percent). If the teacher scored 14 on the free recall test, and their student only scored 7, the student would receive a score of 0.5. Correct free recall scores for all teachers ranged between 0.4 (six correct answers) and 0.93 (14 correct answers), with a mean of 0.63 and a *SD* of 0.14. Adjusted scores for all students ranged between 0.13 and 1.30, with a mean of 0.71 and a *SD* of 0.19. A few students reached scores higher than 1.0 as some teachers may have mentioned items to the students but then failed to list those items themselves on their own recall test. Histograms of the teachers' scores (Fig. 2) and the students' adjusted scores (Fig. 3) are shown above.

Since machine learning allowed us to use a bottom-up approach to capture multiple aspects of gesture, we wanted to cast a wide net for our initial analysis addressing the central question of whether nonverbal behavior could predict learning, but also examine how the behaviors of dyads at various levels of extremity might differ. By removing the middle of the distribution in the machine learning classification, we avoided having to decide that a score of 0.73 was "bad" while a score of 0.75 was "good." Using the distribution from Fig. 3 based on the ratings of success, we divided the pool of participant pairs into several different divisions of high and low. We started with an inclusive definition of high and low scoring pairs, taking the top 27 pairs, with scores above 0.70 and the bottom 23 pairs, with scores below 0.67. These 50 pairs formed our *Inclusive* subset. We then moved to a narrower definition of high and low scoring pairs, using only the top 15 pairs, with scores above 0.80, and bottom 16 pairs, with scores at or below 0.60, to create our *Moderate* subset. Finally, we took only the extremely high and low scoring pairs, comparing the seven participant pairs in which the student had scored 0.92 or more, and the seven pairs with a score of 0.50 or less. These 14 pairs comprised our *Exclusive* subset.

In each instance, we had a nearly equal number of data points in each of the two classes, resulting in a baseline (chance) performance of approximately 50 percent as a comparison point for the success of our classifiers and selected features.

2.5 Nonverbal Feature Extraction

The Kinect output consisted of gesture and posture information from each of the teacher-student dyads time stamps. Then it was labeled according to which Kinect was recording

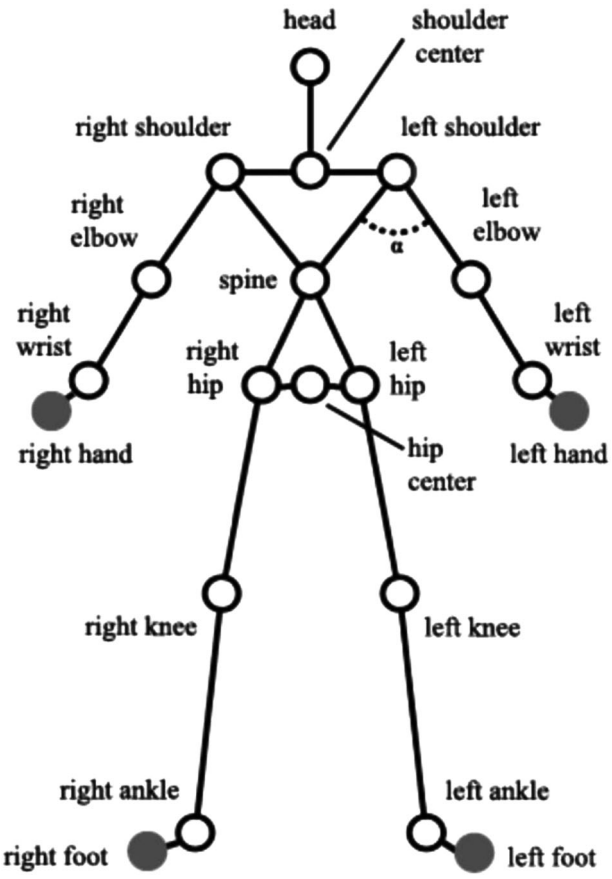

Fig. 4. The modified skeleton derived from Kinect data output in the form of a wireframe. The wireframe consists of 16 nodes with X, Y, and Z values. The nodes in gray were not used in analysis. Angle α describes an example of the angles that we extracted as features. In this case, the angle that represents the shoulder was created from the spine, left shoulder, and left elbow nodes.

the teacher and which was recording the student. Interaction time between teacher and student lasted approximately 3 minutes. In order to ensure that we used exactly the same amount of data from all interactions, we only used the first two minutes of each interaction, as some interactions went longer than others. This provided us with the most consistent data set, as we did not keep pairs for which the first few seconds of data were not available for both participants. Each Kinect recorded at 30 HZ, resulting in approximately 1,800 frames per minute (30 frames per second) for each participant in the interaction. The X, Y, Z positions of the 20 nodes used by the Kinect to represent the joints of the skeleton were recorded, as well as the overall position of the participant in the room calculated by combining those 20 nodes. In addition, for each frame, data was collected on whether each node was tracked, inferred, or not recorded at all.

We used 16 of the 20 nodes to create the modified skeleton seen in Fig. 4. We ignored the averaged node that represented the overall position of the participant, because there was very little variance in this metric due to experimental instructions to the participants to stay on their respective tape marks. We also eliminated four nodes (both hand and both foot nodes), which were not tracked as accurately as the other nodes and were close enough to the wrist and ankle nodes to be fairly redundant. These four deleted features are represented in gray on Fig. 4.

To define our features, we calculated the angles for each Kinect skeleton joint, extracting 18 angles per skeleton (e.g., the angle between the spine-to-left-shoulder "bone" and the left-shoulder to left-elbow "bone"). We were not seeking to identify specific gestures using top-down knowledge of nonverbal communication. Instead, we sought to capture more general qualities of body movement, while staying true to the body's natural anatomy within the confines of the Kinect skeleton. The angles for the movement features were calculated by taking the cosine value of two vectors. An example below shows the distance between the shoulder (S) and elbow (E) nodes (1) and the elbow and wrist (W) nodes (2):

$$Vector(S,E) = (x,y,z)_{elbow} - (x,y,z)_{shoulder} \quad (1)$$

$$Vector(E,W) = (x,y,z)_{wrist} - (x,y,z)_{elbow}. \quad (2)$$

Then, the angle between the two parts was calculated using the distance between the two points, as in the following equation:

$$\theta = arccos(Vector(SE), Vector(E,W)). \quad (3)$$

Some nodes were involved in multiple "joints", resulting in more angles than the number of nodes. For example, the left hip node formed an angle with the left ankle and left knee nodes that roughly corresponded to the movements of the left knee joint. However, the hip joint node was also part of another angle with the spine and left knee that roughly represented the movement of the left hip. All angles used are shown in Table 1.

From each angle we derived a trace of its movement over time, representing the measurement of the angle at each frame, at 30 frames per second. From the changes in this angle from frame to frame we took three measures: mean, standard deviation, and skewness. Thus, for each participant in the dyad, we recorded 54 angle features (three measures for each of 18 angles).

Thus, the mean represented the average angle of the joint, and the standard deviation represented the amount that that angle varied over time. Skewness represented the fact that people may move or bend a joint further (over a wider range) in one direction, and was thus a measure of how much each gesture deviated from the mean, rather than a temporal measure of differing gestures at the beginning or end of an interaction. Thus comparatively large changes in angle would pull the mean away from the mode, skewing it in the direction of these dramatic changes in angle. The formula for skewness is shown below. X represents the variable X, $\boldsymbol{\mu}$ represents the mean, and $\boldsymbol{\sigma}$ represents the standard deviation:

$$\gamma = \sum \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right]. \quad (4)$$

Finally, in order to reduce the redundancy of our feature set and create features that related more clearly to human gestures, we grouped the angle measures into five categories roughly corresponding to the each participant's right

TABLE 1
Movement Features

Features Summed For Head and Torso
1. Angle of head, shoulder center and spine nodes
2. Angle of shoulder center, spine and hip center nodes
3. Angle of right and left shoulder nodes and spine and shoulder center nodes
4. Angle of right and left shoulder nodes, and head and shoulder center nodes
5. Angle of right and left hip nodes, and right and left shoulder nodes
6. Angle of right and left hips and spine and hip center
Features Summed For Right Arm
7. Angle of right and left shoulders and right elbow
8. Angle of spine, right shoulder and right elbow
9. Angle of right shoulder, right elbow and right wrist
Features Summed For Left Arm
10. Angle of left and right shoulders and left elbow
11. Angle of spine, left shoulder and left elbow
12. Angle of left shoulder, left elbow and left wrist
Features Summed For Right Leg
13. Angle of right and left hips and right knee
14. Angle of spine, right hip and right knee
15. Angle of right hip, right knee and right snkle
Features Summed For Left Leg
16. Angle of right and left hips and left knee
17. Angle of spine, left hip and left knee
18. Angle of left hip, left knee and left ankle

arm, left arm, right leg, left leg, and torso/head. Creating sums of mean, standard deviation, and skewness for all the angles between body segments within a body region also allowed us to minimize the effects of occlusion on extremities (e.g., it was unlikely for all joints in a body region to be occluded at once). Each measure represented the entire interaction, such that there was one of each for the entire 2-minute period.

In order to make sure that the movements of the teacher and the student were being recorded during the identical time period, we matched both the beginning and ending timestamps of the north and west Kinects. This ensured that, although our measures were summary measures of the entire time period, they covered exactly the same interval.

2.6 Classification

In order to generalize our results, we compared three different algorithms, a strategy similar to those employed by previous research [13]. To reduce the risk of overfitting and adjust for the fact that our features were not at all independent, we used correlation-based feature selection as a starting point to find the most useful features, and the optimal number of these features to use for our predictions, as described by Salvagnini et al. [27].

Because we had a relatively small data set, we used leave-one-out cross validation as discussed by Witten and Frank [42]. For our outcome measure, we compared the success rates in predicting the students' transformed scores. In order to examine predictive ability between different extremes of successful and unsuccessful teacher and student pairs, we divided the data set into three different cutoff groups, following Baron's general strategy [40]. We also examined which feature sets were most predictive in order to look for meaningful explanations.

In order to be sure that our measures of accuracy were conservative, we strictly separated training and testing data, following the procedure of Castellano et al. [23] and Hoque et al. [13]. Each algorithm was evaluated using leave-one-out cross-validation, removing one dyad in each subset of the data from the original data set to be used as the test data, and all other dyads remain as training data. This was repeated for each dyad in each subset with the test samples removed prior to both feature selection and classifier training, in order to ensure that neither process would overfit to the training data set. In other words, one pair was held out as the test sample prior to feature selection. After feature selection, the number of predictors in the data set was reduced to only the selected features. The resulting model was then tested on the original held-out pair, producing a hit, a miss, a false positive, or a correct rejection. This train/test procedure was repeated n times, until all pairs had been used as the test sample, and the results were summed over the entire interaction. Thus, the pair on which the prediction had been made had not been used for either feature selection or training in that fold.

Following Castellano et al. [23], we used a filter-based method of feature selection; correlation-based feature subset evaluation. In this filter method, individual features are evaluated based on their predictive ability as well as the degree to which they are redundant with other features. This helps to reduce a data set with many features that are not independent.

We selected a decision tree algorithm (J48), in order to help us visualize possible relationships between the data. We selected Multilayer Perceptron (MP) since it is optimized to accurately fit nonlinear patterns in the data. We selected Logistic Regression (LR) because it is a relatively simple but useful classifier to provide a baseline for other classifiers, and it is less likely to overfit than more complex classifiers. Thus, by selecting these three, we tried to create a diverse sample of the available classifiers.

2.7 Tracking Accuracy

The accuracy of the Kinect system compared to other motion capture systems has been tested by other researchers who have found it to be sufficiently accurate to use in naturalistic settings such as the workplace, even if less accurate than other motion capture techniques [43]. However, the format of our experiment may have added extra challenges to tracking. Since the two participants were standing at a conversational distance from one another, some of their gestures may have occluded their conversational partner from their respective cameras. In this case, we would expect lower tracking on the right side of participants who were recorded by the Kinect on the north wall, and lower tracking on the left side of participants who were recorded by the Kinect on the west wall (see Fig. 1). In order to estimate whether or not this occlusion took place, we examined the degree to which nodes were inferred or tracked during the course of the interaction in a random subsample of

TABLE 2
Percentage of Tracked versus Inferred Nodes

Node	North	West
Head	99.85	99.76
Shoulder Center	1	99.99
Left Shoulder	99.41	95.99
Right Shoulder	95.25	99.22
Spine	99.99	99.99
Hip Center	99.99	99.99
Left Elbow	*96.41*	*72.28*
Right Elbow	*73.24*	*99.14*
Left Wrist	*82.21*	*50.99*
Right Wrist	*69.36*	*88.52*
Left Hip	99.99	99.99
Right Hip	99.99	99.99
Left Knee	99.17	97.10
Right Knee	94.38	95.84
Left Ankle	92.58	91.88
Right Ankle	*87.99*	*95.85*

The percentage of nodes tracked as opposed to inferred by the North and West Kinects. Nodes that showed a difference in accuracy between North and West Kinect recordings of greater than 5 percent are italicized. All dyads were counterbalanced, such that teachers were randomly assigned to be recorded by either the North or West Kinect. This was done to ensure that Kinect assignment would not be confounded with participant role.

12 participants (see Table 2). Occlusion was noted in the elbow, wrist and ankle nodes italicized in the table but differences elsewhere in the body were less than 5 percent.

In order to use a conservative measure of accuracy, we only used data that was tracked, and did not include inferred data. In order to be conservative about potential synchrony in our summary measures, for each dyad, we only included time stamps for nodes for which both participants had both nodes tracked (as opposed to inferred). In other words, if the teacher's wrist node was not tracked in a given frame, the student's right wrist node would also be dropped for that time stamp, and none of the angles associated with that node for that time stamp would be calculated.

3 Results

Our goals were to predict high and low success interactions, to begin to assess which features might be driving those interactions, and to find how more inclusive categories of high and low scoring pairs might increase predictive power.

TABLE 3
Predicting Teaching/Learning Success

	Hits	Misses	Correct Rejections	False Positives	Accuracy
Exclusive (14 pairs)					
J48	6	1	6	1	85.7%
MP	6	1	4	3	71.4%
LR	6	1	4	3	71.4%
Moderate (31 pairs)					
J48	6	9	16	0	71.0%
MP	5	10	16	0	67.7%
LR	7	8	10	6	54.8%
Inclusive (50 pairs)					
J48	27	0	0	23	54.0%
MP	21	6	3	20	48.0%
LR	18	9	4	19	44.0%

Inclusive indicates the top 27 and bottom 23 pairs, Moderate the top 15 and bottom 16 pairs, and Exclusive the top 7 and bottom 7 pairs. The classifiers used were decision tree (J48), Multilayer Perceptron (MP) and Logistic Regression (LR).

Table 3 presents the results. Using only a few coarse measures of nonverbal features during a two-minute interaction, we were able to predict whether a teaching/learning interaction would be successful or unsuccessful when looking at more exclusive subsets.

Hits are correct identification of a pair as "good", misses are the incorrect identification of a pair as "bad", correct rejections are the correct identification of a pair as "bad" and false positives are the incorrect identification of a pair as "good". Across all subsets, the J48 Decision Tree algorithm provided the highest degree of accuracy, reaching 85.7 percent for the Exclusive subset. This was significantly above the chance rate of 50 percent ($p < 0.05$). Accuracies were highest for all classifiers used when comparing the most extreme cases of success or failure, with the Exclusive subset of participants. Accuracies declined as the subsets became more inclusive, reaching chance when 50 out of the original 53 pairs were included in the good/bad division.

In order to ensure that overfitting did not occur on the Exclusive data set, we compared leave-one-out cross

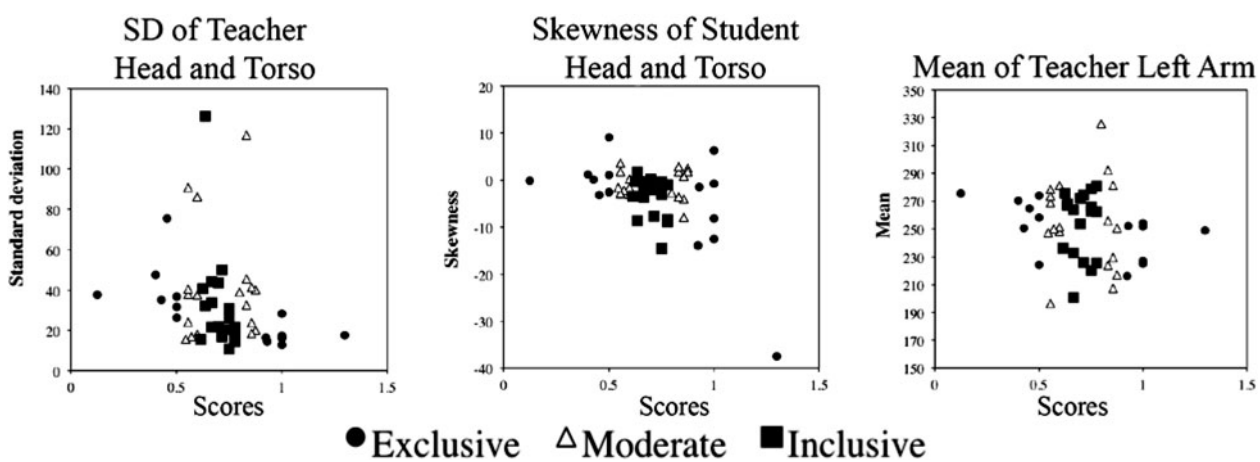

Fig. 5. The three features that demonstrated significant correlations with score are plotted against the scores for each pair.

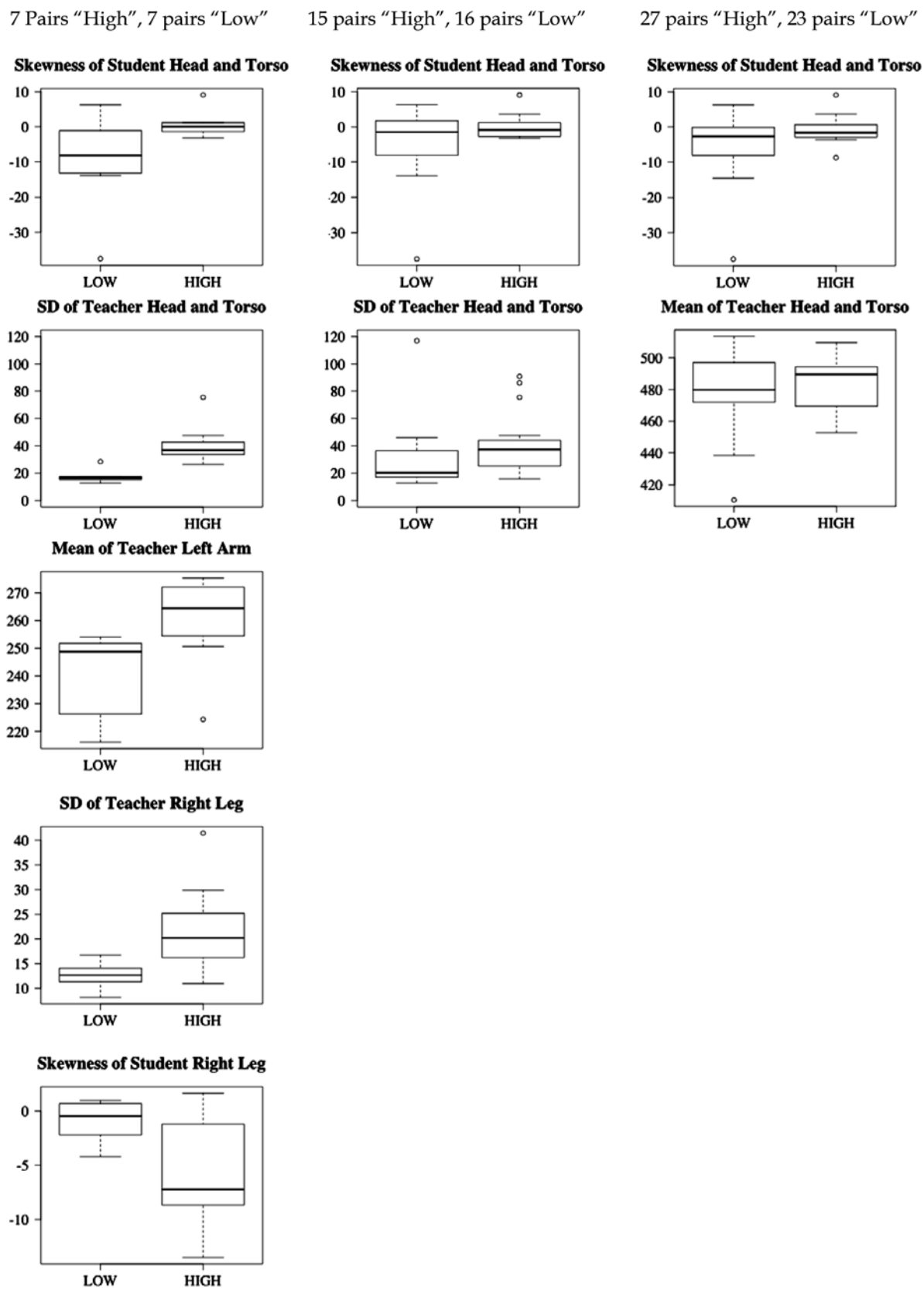

Fig. 6. The above plots show each of the features selected as predictive in at least one fold, for at least one subset, of high and low success pair. Each column represents one subset, beginning with the Exclusive subset on the left hand side. The Y-axis of each plot represents degrees in the case of plots showing the mean or standard deviation (summed across joints within the body region), but is a sum of integral distances from the mean in the case of skewness, which can also be positive as well as negative.

TABLE 4
Percentage of Folds in Which Predictive Features Appear for Each Subgroup

	All Pairs	Inclusive (50 pairs)	Moderate (31 pairs)	Exclusive (14 pairs)
Standard Deviation of Teacher's Head and Torso			44%	100%
Skewness of Student's Head and Torso		12%	56%	20%
Mean of Teacher's Head and Torso		80%		
Mean of Teacher's Left Arm				40%
Standard Deviation of Teacher's Right Leg				32%
Skewness of Student's Right Leg				16%

These percentages reflect the number of folds, averaged across five repetitions, in which these features appear. Features that appeared in fewer than 10 percent of folds are not listed.

validation to five-fold cross-validation, averaging the accuracy over the five repetitions. The patterns of accuracy were similar, with an average accuracy of 74.8 percent over all three algorithms for the five repetitions ($M = 84.3\%$ for J48, $M = 68.6\%$ for MP, $M = 71.4\%$ for LR).

In Fig. 5, the three features that demonstrated significant correlations with score are plotted against the scores for each pair. Fig. 6 shows the different predictive features for each subset. For each subset, between one and five features were chosen for each fold. In order to provide the most predictive features, we used the features that appeared in at least 10 percent of the 25 folds in five-fold cross-validation repeated five times. The distribution of features can be found in Table 4.

In order to learn more about the relationship between features and the outcome, in addition to the machine learning algorithms, we also computed correlations between features and the outcome of the students' adjusted scores. Table 5 presents those results. Summed mean, standard deviation, and skewness measures all appeared as predictive features. The correlations tended to be higher in magnitude for the smaller, more restrictive subsets compared to the larger ones, confirming the pattern that the most extreme success patterns were easiest to predict. The risk of type 1 errors cannot be discounted when dealing with large numbers of correlations. Out of the 24 correlations shown, two were significant at the 0.05 level, two at the 0.01 level, and two at the 0.001 level. Since our feature selection was based on correlation, this is not too surprising, but the magnitude of some of the correlations may provide a starting point for further investigation.

Interpreting these features can be challenging due to the bottom-up nature of how they were computed. The feature comprised of summed standard deviations of the movements of the teacher's head and torso, which was very predictive during machine learning in the most extreme division between high and low success pairs, correlated negatively with the student's transformed score for all subsets, at -0.68. The feature comprised of the summed skewness measures of the student's head and torso showed a similar negative correlation of -0.60. Thus, negative skewness, or movements that decreased the mean of the participants' torso angles, was predictive of outcome. One example of a body movement that might produce negative skewness would be occasional nods in a person who otherwise kept their head more or less upright. This would pull the mean angle of the head over time below the median, upright. However, since neither specific semantic gestures nor directionality were defined in our measure, interpretation remains speculative.

Finally, since synchrony has been indicated as an element of success in teaching and learning, we examined the correlations between the corresponding gestures of the teacher and the student, shown in Table 6. Significant correlations were found in two out of six features. While these correlations may be viewed as predictive features in their own right, examining synchrony will require more granular methods.

4 Conclusion

In the study described above, we demonstrated the ability of an automated affective computing system to analyze naturalistic body movements, and, using these movements, assess the qualities of a teaching/learning interaction. Our results support the view that in such pairs, the nonverbal behavior of both the teacher and student can predict the success of the outcome.

We will first discuss areas that could be improved and then address bigger questions about the direction of future

TABLE 5
Correlations between Predictive Features and Score

	All Pairs	Inclusive (50 pairs)	Moderate (31 pairs)	Exclusive (14 pairs)
Standard Deviation of Teacher's Head and Torso	$-0.26\dagger$	$-0.26\dagger$	-0.28	$\mathbf{-0.68}^{**}$
Skewness of Student's Head and Torso	-0.45^{***}	$\mathbf{-0.48}^{***}$	$\mathbf{-0.51}^{**}$	$\mathbf{-0.60}^{*}$
Mean of Teacher's Head and Torso	0.06	**0.06**	**0.12**	0.38
Mean of Teacher's Left Arm	-0.20	-0.21	$-0.33\dagger$	$\mathbf{-0.54}^{*}$
Standard Deviation of Teacher's Right Leg	-0.18	-0.18	$-.20$	$\mathbf{-0.23}$
Skewness of Student's Right Leg	0.17	0.17	0.27	**0.42**

$\dagger\, p < 0.075, {}^{*}p < 0.05, {}^{**}p < 0.01, {}^{***}p < 0.001$. *Features used in machine learning analysis for each subgroup are in bold print in those columns. (Note that the tests involving fewer pairs are more conservative in terms of significance.)*

TABLE 6
Correlations between Student and Teacher Features

	All Pairs	Inclusive (50 pairs)	Moderate (31 pairs)	Exclusive (14 pairs)
Standard Deviation of Teacher's Head and Torso	0.40**	0.40**	0.33†	**0.31**
Skewness of Student's Head and Torso	−0.03	**−0.04**	**−0.03**	**0.08**
Mean of Teacher's Head and Torso	0.15	**0.16**	**0.28**	0.20
Mean of Teacher's Left Arm	0.13	0.17	0.15	**0.20**
Standard Deviation of Teacher's Right Leg	0.37*	0.36*	0.31†	**0.43**
Skewness of Student's Right Leg	−0.05	−0.06	−0.06	**−0.24**

† $p < 0.075$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *Features used in machine learning analysis for each subgroup are in bold print in those columns. (Note that the tests involving fewer pairs are more conservative in terms of significance.)*

work. Finally, we will discuss how this work intersects with current and future applications of technology detecting affect via body movements.

4.1 Limitations

Our metric for learning, a brief free recall task, is necessarily limited, and does not examine constructivist or active learning. In addition, our sample of convenience, which consisted of university students, may have consisted of more motivated and experienced learners than the average. Further, since both "teacher" and "student" roles were filled by actual student participants, differences in age and authority often apparent in traditional teacher-student roles were not present in our experiment. Finally, in a normal classroom study, the teachers are experts in the subject matter they teach; while in our study the "teachers" had only learned the subject material a few minutes before the students. The correlation between student scores and teacher scores does not take into account differences in interest, enthusiasm and prior knowledge. While teachers with higher scores had more information to impart to students, this relationship in our analysis was complicated by the fact that students' scores were divided by teacher scores, so that students who had very low scoring teachers had to list fewer principles to achieve a "grade" of 100 percent. In addition, teachers may have remembered additional answers during the written test that they did not impart during the teaching interaction, or may have forgotten to write down answers that they had taught to their students. Thus, our metric for capturing learning was imperfect. To truly capture the nature of a teaching interaction, real teachers interacting with real students should be recorded.

In general, the model tended to predict low-scoring pairs more accurately than high-scoring pairs (in other words, there were fewer false positives than misses, overall). This may also be due to our metric of capturing learning. Since success in the interaction was designed to take into account the teacher's score on the material as well, some pairs that had a high adjusted student score might have had their score artificially boosted because the teachers simply did not remember or write down very many environmental principles themselves. Thus, high-scoring pairs included both teachers who learned the material well themselves, and teachers who did not. These pairs may have differed in their behavior.

Other limitations due to the system design must also be considered. The possibility of occlusion when two participants are standing at a conversational distance should not be dismissed, as we see from Table 1 that it was a likely cause of reduced accuracy. However, the tradeoff for environmental validity may be worthwhile for many applications.

Moreover, the requirement that participants be standing, may also well have affected interpersonal dynamics. One area for useful future work is thus to examine how body movements may be predictive in other situations, such as when all participants are seated. As technologies are tailored for specific environments, these limitations may be considered on a case-by-case basis.

The measures that we used for our analysis were fairly coarse and very bottom-up. We did not look at specific gestures, instead summing total movement by body region. Thus, the way in which we derived our features does not allow us to make predictions about the meaning of specific gestures. Also it is important to remember that we cannot make causal inferences from the data- it would be simplistic to conclude that too much movement in the teacher's head and torso leads to poor learning outcomes.

4.2 Future Directions

Following the techniques used by previous work in education, in this initial study we examined subsets of the data consisting of very high and very low scoring pairs. These most extreme cases were predicted with an average accuracy of 76.1 percent, and a high of 85.7 percent, and the magnitude of the linear correlations between predictive features and the accuracy scores generally increased as the cutoff became more restrictive.

While the features chosen using our filter method were primarily drawn from the teachers' movements when predicting the subset of the most extreme division of good and bad, student features were also selected as predictive when examining more general divisions at the Moderate level of 31 pairs. This may indicate that while teacher movements differ greatly in very successful and unsuccessful pairs, student gestures may indicate a certain level of attentiveness that also correlated with success. However, in order to investigate the meaning of these gestures, our teaching and learning task must be refined.

Finally, some of the predictive gestures demonstrated significant correlations between teacher and student gestures. The two features selected to be most predictive, the summed standard deviations of the teacher's head and torso and the summed skewnesses of the student's head and torso seem

likely to be related. Interactional synchrony, perhaps linked with mirroring or matching behavior, may have been a component of the predictive power of the interaction, although the lack of a temporal component to our feature set makes it difficult to prove. Interpersonal synchrony has been studied in many contexts (for a review see Delaherche et al. [44]). Evidence has been found that manipulating synchrony can increase rapport [45], and in two studies particularly relevant to learning, induced interpersonal synchrony can increase memory of the partner's speech [46], [47]. The more specific synchronous behavior of mimicry appears similarly important in both generating and recognizing rapport and other affiliative behaviors (for a review, see Chartrand and Lakin [48]). Adding ratings of rapport and correlating these ratings with a quantitative measure could provide additional support to a hypothesis of synchrony supported by a more fine-grained analysis of gesture.

Although it is tempting to over-interpret the features that are selected to be most significant, it is important to keep in mind that the set of features selected relies on relationships within the models that are not necessarily intuitive. Further investigation is required to interpret these findings.

Finally, while we examined the predictive power of features from both the teacher and the student, other features, such as proximity, were not included in our model. Such combined features may hold the potential of being very predictive, though of course research would need to take into account differences in culture, gender, age, degree of acquaintanceship and other individual differences. These features are natural targets for future work in this area. As suggested by other research [49], adding higher level, semantically interpretable features may improve accuracy. In addition, beyond body movement, adding the tracking of other modalities is also likely to lead to greater success in affect detection [50].

Accuracies of 85.7 percent were obtainable using only 120 seconds of the interaction. This aligns with previous research demonstrating the effectiveness of a thin slice of observation in interpreting interactions [28]. Determining whether behavior at the beginning, middle, or end of an interaction is most predictive is a potentially productive area of investigation.

Our ability to predict outcomes also speaks to the potential of active computer vision systems, such as the Kinect, to collect useful data in a naturalistic environment. Although a degree of occlusion took place, as evidenced by the percentages of nodes that were inferred rather than tracked depending on the camera angle (see Table 1), our algorithms were still able to make predictions at rates considerably higher than chance. Although the skeleton derived from Kinect tracking data is clearly not completely anatomically accurate, we were able to make useful predictions without even using all of the available nodes. This underlines the usefulness of low-level features in general affect detection that is not specific to a given situation. It also implies the opportunities that may exist for collecting data through other interfaces, such as touchscreens [25].

Future research may include comparing changes in a participant's nonverbal behavior, and changes in outcome, with different interaction partners. Another avenue could be to compare nonverbal behavior and outcomes over time, which might provide clues to how interactions can be guided for better outcomes.

Another fruitful area of investigation may be considering intercultural communication through the lens of gestural interaction. The presentation and interpretation of affect from body posture has also been shown to change depending on the culture of the observer [51]. Do the simple gestural measures we obtained from a group of American undergraduates apply cross-culturally? Can these measures be used to improve cross-cultural communication, or can other tools be built to assess these kinds of interactions?

4.3 Possible Applications

Embodiment in interactive environments is increasingly validated as important to engagement, social interactions, and enjoyment. Learning what nonverbal behavior to track, and how it should most effectively be rendered in a mediated or virtual environment is important in designing and assessing actions in such environments. Extensive and ongoing research has examined what kinds of nonverbal behavior embodied agents should utilize for greatest effectiveness [52], including how this behavior should be guided by the nonverbal behavior of the human conversational partner. Further examining what movements may be most helpful to further an interaction will aid in these goals.

Optimizing partnerships in general, either by assessing nonverbal behavior in real life, or mediating nonverbal behavior effectively, is one very interesting possible arena for applications. Giving people feedback on the nonverbal components of their interaction in real time may allow them to adjust their nonverbal behavior to positively affect the outcome of those interactions. For example, providing teachers or tutors with this kind of feedback in real time could improve teaching outcomes. Physicians interacting with patients might use this to practice building rapport with patients. Leveraging the tracking of nonverbal behavior can even be a tool to learn to reduce social anxiety by improving social skills (similar to recent work utilizing facial expression [53]), or to aid in conflict resolution. In addition, such tracking of nonverbal behavior could be used to improve the nonverbal behavior of embodied agents.

Beyond the general importance of rapport generated by nonverbal communication, the success of teaching and learning situations in particular may involve body movements. Embodied cognition researchers propose that information processing is conducted using the body [54]. This view is supported by work by Goldin-Meadow and colleagues, which indicates that gestures may signal important stages in learning [55], and that the way teachers recognize and react to these gestures may help to determine learning outcomes [56].

In addition, gestures and body movement in general can also change the person who engages in them, physiologically, psychologically and behaviorally. This means that detecting and offering feedback on gestures and body movements can be leveraged to good effect in a number of areas. Because tracking body movements in particular may reveal behavior of which the participants themselves may not be aware, such systems may also hold the possibility of

providing information that can assist an interaction in real time. For example, feedback on body movements has been suggested as a method to mitigate chronic pain [57]. Encouraging body movement has been proposed as crucial for game applications in particular [58], especially in increasing engagement, enjoyment, and affective experiences [59], [60]. Increasing the extent to which gestures can be tracked and incorporated into games has been proposed to decrease anxiety in movement based learning games [61] and increase social engagement in collaborative games [62], [63].

Ethical concerns may well arise with the use of this technology to assess individual performance [64]. While assessing the quality of interpersonal interactions has obvious applications, both users and developers must be mindful that the measurements taken do not necessarily reflect the qualities of the individuals involved, but whether a single short-term interaction between two individuals in a dyad is likely to be successful. Thus, this technology may be more usefully applied to optimize partnerships, for example, by providing feedback to existing dyads, or by reassigning individuals to pairs whose nonverbal behavior predicts better learning. In addition, a balance must be struck between recording behavior without participants' awareness, risking deception, and making participants self-conscious about being monitored, which might reduce the validity of the predictions as well as create an oppressive environment.

The automatic assessment of gesture can, thus, not only predict behavior, but also may provide users with new tools to understand and engage with their own behavior in ways that have never before been possible. As D'Mello and Calvo point out, [65] more objective methods of data collection and analysis can guide the development of new technologies, as well as promote the study of affect's impact on activities in the real world.

Acknowledgments

The work presented herein was funded in part by Konica Minolta as part of a Stanford Media-X grant. In addition, this research was supported by funding from the National Science Foundation–Grant #0966838. The conclusions reached are those of the investigators and do not necessarily represent the perspectives of the funder. Funding was also provided by the Dutch national program COMMIT. The authors thank Konica Minolta for the valuable insights provided by their visiting researchers, especially Dr. Haisong Gu. They thank the staff of the Stanford Virtual Human Interaction Lab (VHIL), especially lab manager Cody Karutz, as well as Jimmy Lee, Pam Martinez, Evan Shieh, Le Yu, Suzanne C. Stathatos, and Alex Zamoschin for their programming assistance. Finally, we thank RISE interns Karena Chicas and Rocio Linares for organizing the data, and Brian Perone, Clare Purvis, Erik Brockbank, David Groff and Howon Lee for their helpful comments.

References

[1] A. M. Kessell and B. Tversky, "Gestures for thinking and explaining," in *Proc. Cogn. Sci. Soc. Meetings*, 2005, p. 2498.

[2] W. Roth, "Gestures: Their role in teaching and learning," *Edu. Res.*, vol. 71, no. 3, pp. 365–392, 2011.

[3] S. Mota and R. W. Picard, "Automated posture analysis for detecting learner's interest level," in *Proc. IEEE Conf. Comp. Vision Pattern Recog. Workshop*, 2003, vol. 5, p. 49.

[4] T. Dragon, I. Arroyo, B. Woolf, W. Burleson, R. el Kaliouby, and H. Eydgahi, "Viewing student affect and learning through classroom observation and physical sensors," in *Intelligent Tutoring System*. Berlin, Germany: Springer, 2008, pp. 29–39.

[5] S. D'Mello, T. Jackson, S. Craig, B. Morgan, P. Chipman, H. White, N. Person, B. Kort, R. el Kaliouby, R. W. Picard, and A. Graesser, "AutoTutor detects and responds to learners affective and cognitive states," in *Proc. Workshop Emotional. Cogn. Issues Intell. Tutoring Syst. Conjunction 9th Int. Conf. Intell. Tutoring Syst.*, 2008, pp. 31–43.

[6] S. K. D'Mello and A. Graesser, "Automatic detection of learner's affect from gross body language," *Appl. Artif. Intell.*, vol. 23, no. 2, pp. 123–150, 2009.

[7] M. LaFrance and M. Broadbent, "Group rapport: Posture sharing as a nonverbal indicator," *Group Organization Stud.*, vol. 1, pp. 328–333, 1976.

[8] F. J. Bernieri, "Coordinated movement and rapport in teacher student interactions," *J. Nonverbal Behav.*, vol. 12, pp 120–138, 1988.

[9] J. Zhou, "The effects of reciprocal imitation on teacher-student relationships and student learning outcomes," *Mind, Brain, Edu.*, vol. 6, no. 2, pp. 66–73, 2012.

[10] T. O. Meservy, M. L. Jensen, J. Kruse, J. K. Burgoon, and F. Jay, "Detecting deception through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 36–42, Sep./Oct. 2008.

[11] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling naturalistic affective states via facial, vocal and bodily expressions recognition," in *Proc. 8th Int. Conf. Artif. Intell. Human Comput.*, 2007, pp. 92–116.

[12] M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *J. Multimedia*, vol. 1, no. 6, pp. 22–35, 2005.

[13] M. E. Hoque, D. J. McDuff, and R. W. Picard, "Exploring temporal patterns in classifying frustrated and delighted smiles," *IEEE Trans. Affect. Comput.* , vol. 3, no. 3, pp. 323–334, Jul.–Sep. 2012.

[14] M. E. Jabon, S. J. Ahn, and J. N. Bailenson, "Automatically analyzing facial-feature movements to identify human errors," *IEEE Intell. Syst.*, vol. 26, no. 2, pp. 54–63, Mar./Apr. 2011.

[15] J. H. Janssen, P. Tacken, J. J. G. DeVries, E. L. Van den Broek, J. H. M. Westerink, P. Haselager, and W. A. IJsselstein, "Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection," *Human-Comput. Interact.*, vol. 28, no. 6, pp. 479–517, 2012.

[16] A. Kleinsmith, and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 1–31, Jan./Mar. 2013.

[17] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: State-of-the-art and future perspectives of an emerging domain," in *Proc. 16th ACM Int. Conf. Multimedia*, 2008, pp. 1061–1070.

[18] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image Vis. Comput.*, vol. 31, no. 2, pp. 137–152, 2012.

[19] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[20] A. Jamalian and B. Tversky. "Gestures alter thinking about time," in *Proc. 34th Annu. Conf. Cogn. Sci. Soc.*, 2012, pp. 551–557.

[21] D. R. Carney, A. J. C. Cuddy, and A. J. Yap, "Power posing brief nonverbal displays affect neuroendocrine levels and risk tolerance," *Psychol. Sci.*, vol. 21, no. 10, pp. 1363–1368, 2010.

[22] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. Driessen, "Gesture-based affective computing on motion capture data," in *Proc. 1st Int. Conf. Affect. Comput. Intell. Interact.*, 2005, pp. 1–7.

[23] G. Castellano, S. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. 2nd Int. Conf. Affect. Comput. Intell. Interact.*, 2007, pp. 71–82.

[24] J. N. Bailenson, N. Yee, S. Brave, D. Merget, and D. Koslow, "Virtual interpersonal touch: Expressing and recognizing emotions through haptic devices," *Human-Comput. Interact.*, vol. 22, no. 3, pp. 325–353, 2007.

[25] Y. Gao, N. Bianchi-Berthouze, and H. Meng, "What does touch tell us about emotions in touchscreen-based gameplay?" *ACM Trans. Comput.-Human Interact.*, vol. 19, no. 4, p. 31, 2012.
[26] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 4, pp. 1027–1038, Jan. 2011.
[27] P. Salvagnini, H. Salamin, M. Cristani, A. Vinciarelli, and V. Murino, "Learning how to teach from video lecture: Automatic prediction of lecture ratings based on teacher's nonverbal behavior," in *Proc. IEEE 3rd Int. Conf. Cogn. Infocommun.*, 2012, pp. 415–419.
[28] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Towards a histology of social behavior: Judgmental accuracy from thin slices of behavior," *Adv. Exp. Soc. Psychol.*, vol. 32, pp. 201–272, 2000.
[29] N. Ambady, D. Laplante, T. Nguyen, R. Rosenthal, N. Chaumeton, and W. Levinson, "Surgeons' tone of voice: A clue to malpractice history," *Surgery*, vol. 132, pp. 5–9, 2002.
[30] E. Babad, F. Bernieri, and R. Rosenthal, "Nonverbal communication and leakage in the behavior of biased and unbiased teachers," *J. Personality Soc. Psychol.*, vol. 56, pp. 89–94, 1989.
[31] F. Ramseyer and W. Tschacher, "Nonverbal synchrony in psychotherapy: Coordinated body-movement reflects relationship quality and outcome," *J. Consulting Clin. Psychol.*, vol. 79, no. 3, pp. 284–295, 2011.
[32] G. Johansson, "Visual perception of biological motion and a model for its analysis' attention," *Perception Psychophys.*, vol. 14, no. 2, pp. 201–211, 1973.
[33] G. Mather and L. Murdoch, "Gender discrimination in biological motion displays based on dynamic cues," *Proc.: Biol. Sci.*, vol. 258, no. 1353, pp. 273–279, 1994.
[34] K. L. Johnson, S. Gill, V. Reichman, and L. G. Tassinary, "Swagger, sway, and sexuality: Judging sexual orientation from body motion and morphology," *J. Personality Soc. Psychol.*, vol. 93, pp. 321–334, 2007.
[35] J. Michalak, N. F. Troje, and T. Heidenreich, "The effects of mindfulness-based cognitive therapy on depressive gait patterns," *J. Cogn. Behav. Psychotherapies*, vol. 11, pp. 13–27, 2011.
[36] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, no. 6, pp. 717–746, 2004.
[37] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vis. Image Understanding*, vol. 108, pp. 4–18, 2007.
[38] C. C. Martin, D. C. Burkert, K. R. Choi, N. B. Wieczorek, P. M. McGregor, R. A. Herrmann, and P. A. Beling, "A real-time ergonomic monitoring system using the Microsoft Kinect," in *Proc. IEEE Syst. Inf. Des. Symp.*, 2012, pp. 50–55.
[39] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from RGBD images," *Plan, Activity, and Intent Recognition* p. 64, 2011.
[40] B. Barron, "When smart groups fail," *J. Learn. Sci.*, vol. 12, no. 3, pp. 307–359, 2003.
[41] Microsoft Corp. Redmond WA, USA. Kinect for Xbox 360, 2011.
[42] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2005.
[43] T. Dutta, "Evaluation of the Kinect^{TM} sensor for 3-D kinematic measurement in the workplace," *Appl. Ergonomics*, vol. 43, no. 4, pp. 645–649, 2012.
[44] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Trans. Affect. Comput.* , vol. 3, no. 3, pp. 349–365, Jul.–Sep. 2012.
[45] S. S. Wiltermuth and C. Heath, "Synchrony and cooperation," *Psychol. Sci.*, vol. 20, no. 1, pp. 1–5, 2009.
[46] L. K. Miles, L. K. Nind, Z. Henderson, and C. N. Macrae, "Moving memories: Behavioral synchrony and memory for self and others," *J. Exp. Soc. Psychol.*, vol. 46, no. 2, pp. 457–460, 2010.
[47] C. N. Macrae, C. Neil, O. K. Duffy, L. K. Miles, and J. Lawrence, "A case of hand waving: Action synchrony and person perception," *Cognition*, vol. 109, no. 1, pp. 152–156, 2008.
[48] T. L. Chartrand and J. L. Lakin, "The antecedents and consequences of human behavioral mimicry," *Annu. Rev. Psychol.*, vol. 64, pp. 285–308, 2013.
[49] A. Kleinsmith, T. Fushimi, N. Bianchi-Berthouze, "An incremental and interactive affective posture recognition system," in *Proc. Int. Workshop Adapt. Interact. Style Affect. Factors, Conjunction Int. Conf. User Model.*, 2005, pp. 1–13.
[50] S. D'Mello and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Model. User-Adapt. Interact.*, vol. 20, no. 2, pp. 147–187, 2010.
[51] A. Kleinsmith, P. Ravindra De Silva, and N. Bianchi-Berthouze, "Cross-cultural differences in recognizing affect from body posture," *Interact. Comput.*, vol. 18, no. 6, pp 1371–1389, 2006.
[52] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy, "Creating rapport with virtual agents," in *Intelligent Virtual Agents*. Berlin, Germany: Springer, 2007, pp. 125–138.
[53] R. El. Kaliouby, A. Teeters, and R. W. Picard, "An exploratory social-emotional prosthetic for autism spectrum disorders," in *Proc. Wearable Implantable Body Sens. Netw., Int. Workshop*, 2006, pp. 3–4.
[54] P. M. Niedenthal, L. W. Barsalou, P. Winkielman, S. Krauth-Gruber, and F. Ric, "Embodiment in attitudes, social perception, and emotion," *Pers. Soc. Psychol. Rev.*, vol. 9, no. 3, pp. 184–211, 2005.
[55] S. Goldin-Meadow and M. A. Singer. "From children's hands to adults' ears: Gesture's role in the learning process," *Develop. Psychol.*, vol. 39, no. 3, pp. 509–519, 2003.
[56] S. Goldin-Meadow and S. M. Wagner, "How our hands help us learn," *Trends Cogn. Sci.*, vol. 9, no. 5, pp. 234–241, 2005.
[57] M. S. H. Aung, B. Romera-Paredes, A. Singh, S. Lim, N. Kanakam, A. C. de C. Williams, and N. Bianchi-Berthouze, "Getting rid of pain-related behavior to improve social and self-perception: A technology-based perspective," in *Proc. 14th Int. Workshop Image Anal. Multimedia Interact. Serv.*, 2013, pp. 1–4.
[58] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," *Human-Comput. Interact.*, vol. 28, no. 1, pp. 40–75, 2013.
[59] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience," *IEEE Trans. Comput. Intell. Artif. Intell. Games*, vol. 4, no. 3, pp. 199–212, Sep. 2012.
[60] N. Bianchi-Berthouze, P. Cairns, A. Cox, C. Jennett, and W. W. Kim, "On posture as a modality for expressing and recognizing emotions," in *Proc. Emotion HCI*, 2008, pp. 74–80.
[61] K. Isbister, M. Karlesky, and J. Frye, "Scoop! Using movement to reduce math anxiety and affect confidence," in *Proc. Int. Conf. Found. Dig. Games*, 2012, pp. 228–230.
[62] K. Isbister. "How to stop being a buzzkill: Designing yamove!, a mobile tech mash-up to truly augment social play," in *Proc. 14th Int. Conf. Human-comput. Interact. Mobile Devices Serv. Companion*, Sep. 2012, pp. 1–4.
[63] S. E. Lindley, J. Le. Couteur, and N. L. Berthouze, "Stirring up experience through movement in game play: Effects on engagement and social behaviour," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2008, pp. 511–514.
[64] R. W. Picard, "Affective computing: Challenges," *Int. J. Human-Comput. Stud.*, vol. 59, no. 1, pp. 55–64, 2003.
[65] R. A. Calvo and S. K. D'Mello, *New Perspectives on Affect and Learning Technologies*, vol. 3. New York, NY, USA: Springer, 2011.

Andrea Stevenson Won received the MS degree in biomedical visualization from the University of Illinois, Chicago, in 2005. She is currently working toward the PhD degree at the Department of Communication, Stanford University. Her research interests include mediated self-representation and capturing, assessing and manipulating body movements to affect outcomes. She is a member of the International Communication Association.

Jeremy N. Bailenson is an associate professor at Stanford's Department of Communication.

Joris H. Janssen received the MSc (cum laude) degree in artificial intelligence from the Radboud University, Nijmegen in 2008, and the PhD (cum laude) degree in human technology interaction from the Eindhoven University of Technology in 2012. He is a senior researcher at Sense Observation Systems and a Media-X visiting scholar at Stanford Univeristy. His research interests include affective computing, social signal processing, persuasive technology, health psychology, human-computer interaction, and transformed social interaction.