J Nonverbal Behav
DOI 10.1007/s10919-014-0186-0

ORIGINAL PAPER

Automatically Detected Nonverbal Behavior Predicts Creativity in Collaborating Dyads

Andrea Stevenson Won · Jeremy N. Bailenson ·
Suzanne C. Stathatos · Wenqing Dai

Abstract In the current study we administered a creative task in which two people collaboratively generated novel strategies to conserve resources. During this task, the nonverbal behavior of 104 participants in 52 pairs was tracked and recorded using the Kinect computer vision algorithm. We created a measure of synchrony by correlating movements between the two dyad members, and showed that synchrony occurred—that is, correlations decreased when we increased delay between the recorded movements of pair members. We also demonstrated a link between nonverbal synchrony and creativity, as operationalized by the number of new, valid ideas produced. Linear correlations demonstrated a significant relationship between synchrony and creativity. Finally, models using synchrony scores as input predicted whether dyads were high or low in creativity with a success rate as high as 86.7 % in the more exclusive subsets. We discuss implications for methodological approaches to measuring nonverbal behavior and synchrony, and suggest practical applications which can leverage the current findings.

Keywords Nonverbal behavior · Synchrony · Gesture · Collaboration · Creativity · Kinect · Interpersonal communication · Contingency

Introduction

Understanding creativity in groups is both theoretically interesting and has a wealth of practical applications. Some research indicates that pairs can be more creative than individuals working alone (Torrance 1970), while other work has investigated which factors contribute to creativity in groups (Kurtzberg and Amabile 2001), including diversity (McLeod et al. 1996), individualism (Goncalo and Staw 2006), and mood (Baas et al.

A. S. Won (✉) · J. N. Bailenson · S. C. Stathatos · W. Dai
Department of Communication, Stanford University, 450 Serra Mall, Stanford, CA 94305-2050, USA
e-mail: aswon@stanford.edu

Published online: 22 May 2014

2008). However, there is little quantitative data that examines the relationship between nonverbal behavior and creativity.

The concept of rapport, “a state of mutual positivity and interest that arises through the convergence of nonverbal expressive behavior in an interaction” (Drolet and Morris 2000, p. 27), has been linked to success in a number of interpersonal interactions, including physician/patient interactions (Harrigan et al. 1985) and conflict resolution (Drolet and Morris 2000). In developing virtual agents, in which the appearance and action of the agents are designed to facilitate a simulated interpersonal interaction, rapport is also an important metric (Huang et al. 2011) by which the success of the agent is judged. Thus, a metric of rapport may possibly also be used to predict the outcome of an interaction.

In one of the first empirical demonstrations of this phenomenon, La France and Broadbent (1976) found that the synchronous quality of posture changes between a teacher and students in a classroom setting was linked to student reports of classroom rapport. This so-called *synchrony* (the temporal linkage of the nonverbal behavior of two or more interacting individuals) has been connected to rapport by multiple researchers (e.g., Bernieri 1988; Tickle-Degnen and Rosenthal 1990). The concept of synchronous nonverbal behavior was first introduced by Condon and Ogston (1966) who filmed and compared the flow of conversations with normal, aphasic, and schizophrenic participants. This was followed by Kendon (1970), who used films of participants’ informal conversations to examine synchronous behavior. Kendon proposed that “Coordination of movement in interaction may thus be of great importance since it provides one of the ways in which two people signal that they are ‘open’ to one another, and not to others…indicating or promoting rapport in a number of cases” (p. 124). The possible link between synchrony and behavior matching has been further investigated as indicating or promoting rapport. For example, the success of patient/therapist relationships (Ramseyer and Tschacher 2011) was predicted by synchronous gesture. Thus, we see not only that synchronous behavior may be linked to the concept of rapport, but that it is linked to a number of other positive outcomes in interpersonal interactions (Pentland 2010).

Assessing the synchrony or perceived rapport of interacting individuals is both easy and difficult. Humans have a natural ability to recognize qualities, like synchrony, in others’ interactions. This was demonstrated by Bernieri (1988), who tested coders’ ability to identify synchrony at levels above chance. He did this by comparing the ratings of filmed clips of teacher/student interactions to “pseudo-interactions… in which the recorded behavior of people in two different interaction dyads were combined to appear as though they were interacting with each other” (p. 123). Actual interactions received higher ratings than pseudo interactions, indicating that human coders were indeed able to perceive synchrony.

Quantifying synchrony, however, or coding any other kind of nonverbal behavior using human coders, can be problematic. Coders must either covertly observe participants, or, more often, behavior is recorded and then painstakingly coded frame by frame. This process is expensive and slow; consequently sample sizes from synchrony studies tend to be small. In addition, human coders bring their preconceived notions to bear on the task, possibly introducing bias. For example, in an experiment that used the same video recording of two hands moving in synchrony, but varied only the apparent skin color of each pair, participants gave lower ratings of synchrony to pairs that appeared to be composed of participants with different skin colors (Lumsden et al. 2012). In order to focus on specific qualities of the interaction, for example, temporal synchrony, other information may actually have to be removed from the data in post-processing so that coders are not influenced by information irrelevant to the research question. In a 1994 analysis by

Bernieri et al., participants' faces were blurred in the video and audio was removed, so that coders were not distracted by these elements. In a 1999 experiment by Grahe and Bernieri, observers did not improve in their judgments of rapport when they were given access to verbal content, and verbal content "may even have worsened" (p. 264) their judgment.

Automating the ability to interpret body movements can also vastly increase the amount of data that can be processed and interpreted, increasing researchers' ability to address questions about the links between body movement, rapport, and interaction outcome. Thus, researchers' attention has turned to various methods of automatically detecting and interpreting body movements. For example, looking at more general measures of body movement has also promoted understanding of how nonverbal channels might indicate deception (Meservy et al. 2005) or reveal affect (Kleinsmith and Bianchi-Berthouze 2007).

However, automating the assessment of body movement requires the ability to "see" body movements. This problem has been addressed using various kinds of computer vision. These methods include very complex technological setups, for example, placing optical markers on participants' joints (Kapur et al. 2005) as well as less expensive and more portable methods such as the summation of pixels from video (Schmidt et al. 2012; Ramseyer and Tschacher 2011). All of these methods have their pros and cons. While marker based systems can be extremely accurate, they are also expensive and can be cumbersome and distracting. While video-based techniques are unobtrusive and inexpensive, the measures that can be drawn from them are often general and poor lighting, occlusion, or bad camera angles can lessen their value.

Recent advances in *active* computer vision have begun to address some of these problems, resulting in systems designed for commercial use that are cheap, portable, and noninvasive, and do not require the use of markers. These systems, including the Microsoft Kinect, which uses an infrared emitter and sensor, use algorithms to generate skeleton models of the human user in real time, and thus provide a compromise between maximizing accuracy and unobtrusive, natural data collection. Such systems have been increasingly used by researchers to detect human gestures in a naturalistic environment (Martin et al. 2012; Sung et al. 2011).

In order to process the enormous quantities of data produced by such a system, machine learning interfaces (e.g.,Witten et al. 2011) can be used to derive bottom-up algorithms that predict outcome measures. In the realm of nonverbal behavior, this strategy has been employed in studies using facial feature movements to predict error (Jabon et al. 2011a) or unsafe driver behavior (Jabon et al. 2011b). Similar methods have been used to infer states of mind. Castellano et al. (2007) and Kapur et al. (2005) used machine learning techniques combined with computer vision to attempt to differentiate between body movements associated with various enacted emotions, while Hoque et al. (2012) investigated the distinction between frustrated versus happy smiles. The advantage of this kind of analysis is that although machines do not have the natural human abilities to identify and interpret gestures, they also bring no preconceived notions to the task. Their ability to process vast quantities of data provides researchers with a valuable tool to assess large quantities of data and look for new patterns in the results.

In a previous study (Won et al. 2014), movement data from a set of 53 teaching and learning pairs was analyzed combining computer vision and machine learning. Microsoft Kinect was used to capture the movements of two participants engaged in a verbal teaching/learning task. The Kinect's computer vision algorithm creates a "skeleton" with 20 nodes that approximate the major joints of the body, and the positions of these nodes in X Y Z space were recorded at a rate of 30 frames per second. Using these data, features were created from the movement of these points over time, which captured aspects of the movements of the participants over the course of the interaction. These features were then used as input in machine learning algorithms to predict the outcome of the interaction, which was classified as successful or unsuccessful based on the participants' free recall scores post-interaction. In the current study, we examined a unique dataset that consists of a follow-up task using a subset of the participants from the previous study. We leveraged the substantive information they gained in the pedagogical task as a springboard to engage in the creativity collaboration reported in the current paper.

We chose a creative collaborative task as a contrast to the less balanced interaction of teaching/learning (which retains a power dynamic between teacher and student). Creativity collaborations are thought to uniquely leverage group synergy (Kurtzberg and Amabile 2001).

Hence a creative task lent itself to finding nonverbal correlates from both interactants, where both had the opportunity to contribute equally to the outcome of the interaction.

In the current study, we analyzed the gestures of 104 individuals from 52 participant pairs. We first examined the data for evidence of synchrony between participants. We then used synchrony to predict each pair's creativity score. By recording unprompted movements and open responses in an unscripted creative task, we hoped to capture nonverbal behavior in a dyadic interaction in a naturalistic manner. Transcribing and rating the responses to the task provided a measure of creative success for each dyad.

Method

Participant Population

A sample of 138 participants was drawn from the student population and experiment participant list of a medium-sized North American university. Thirty-four were removed due to equipment failure or overly short recording times, leaving 51 male and 53 female participants, between the ages of 18 and 22. All were randomly assigned to pairs, resulting in 18 female–female, 13 male–male, and 21 mixed-gender pairs. Participants received either course credit or a $15 gift card for their participation, and all signed an informed consent form before beginning any part of the experiment.

Apparatus

Both participants were recorded by a Kinect camera, mounted on the wall approximately 1.5 m in front and .5 m to the left of them. Figure 1 provides a bird's eye view of the scene, and Fig. 2 shows a photograph of a sample participant pair. The Kinect also recorded audio during the entire interaction, which was then transcribed and coded by research assistants to create the outcome measures.

Procedure

For our interaction, we selected a collaborative creative task in which both participants would have an opportunity to contribute equally to the outcome. We prepared each participant pair by having them complete a learning task involving memorizing and recalling 15 environmental principles related to water use (as described in Won et al. 2014).

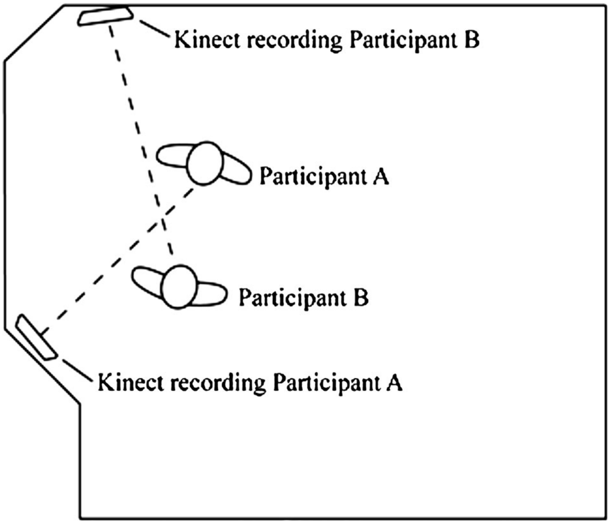

Kinect recording Participant B
Participant A
Participant B
Kinect recording Participant A

Fig. 1 This bird's eye view of the room shows the relationship of the participants to the Kinects recording them

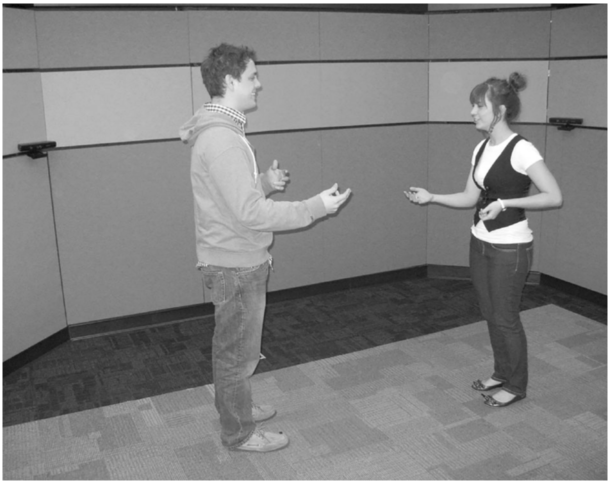


Fig. 2 The Kinects can be seen on the walls behind the sample participant pair in the experiment room

As can be seen in Fig. 2, the Kinects were clearly visible on the walls. However, the main experiment room, also used for virtual reality experiments, contained a number of other cameras and tracking devices so that the Kinects were not the focus of participants' attention. Participants were brought into the main experiment room, asked to stand on tape marks and told that in order for the audio to be recorded they should try to keep their positions. The audio recording was indicated by gesture as coming from the ceiling, in order to direct participants' attention away from the wall-mounted Kinects. It should be noted that while this requirement kept participants in view of the sensors, it also limited the extent to which interpersonal distance could be observed, since participants' positions relative to each other were essentially fixed. The positions of participants' tape marks were carefully piloted to make sure that the Kinects accurately tracked the participants.

The experimenter told the participant pairs that, as soon as she left the room, they would have 5 min to generate as many "new, good" ideas dealing with conserving water or energy or reducing water or energy use, as they could. Participants were told that their responses would be recorded, so they did not need to remember anything or write anything down. Recording of the participants' gestures began as soon as the door closed behind the experimenter. After 5 min, the experimenter re-entered the room and informed participants that they had completed the task. Participants then filled out a brief demographic questionnaire and the experiment was over.

Measures

Creativity

The audio files were transcribed and then the transcripts were scored for creativity by two independent coders. *Creativity* was measured following Oppezzo and Schwartz (2014) who operationalized creativity as "appropriate novelty" using criteria modified from Guilford (1957). Thus, novel ideas were ones that differed from a set of provided prompts and were appropriate to the topic at hand. For this measure, we summed every unique idea and gave it an initial value of 1. Ideas that were clearly inappropriate or facetious (i.e., "stop drinking water") were reduced to a score of zero and thus did not contribute to the final score. This measure, of *quantity* of valid ideas, essentially scored the fluency of a pair's idea generation, which is a common metric to most creativity tests. We chose this measure in order to compare it to a possible confound, talkativeness, which might also be reflected in the nonverbal behavior of the participants. A second measure, *quality* of ideas, gave an additional point to ideas that deviated from provided prompts. For example, if participants were taught to reduce shower time by using a timer and changed the wording such that it became an idea about using a timer to save electricity, they would receive one for a quantity score but zero for a quality score. On the other hand, if they diverged from the provided strategy both in domain and in method, for example, using social media to reduce car use, they would receive one for a quantity score and one for a quality score We calculated two separate scores for each participant; one for quantity (the number of appropriate ideas) and an additional score for quality (the number of ideas that received an additional point for "appropriate novelty"), and summed them for our final creativity measure for each pair.

For quantity, ratings of the number of appropriate ideas by each coder correlated at .87, so we averaged the rating into a single pair score. Pair scores ranged between 4.5 and 21.0 ($M = 10.9$, $SD = 3.7$).

For quality, ratings of the number of divergent ideas by each coder correlated at .78, so we averaged the ratings into a single pair score. Quality pair scores ranged from zero to 14.5 ($M = 4.7$, $SD = 2.8$).

To compute the overall *creativity* scores, we summed the quantity and quality scores. Ratings of the summed creativity scores correlated between raters at .87, so we averaged across the two coders. Creativity scores ranged from 4.5 to 35.5 ($M = 15.6$, $SD = 6.1$). Figure 3 shows score distribution. Table 1 shows the relationship between these measures, as well as the measure of word count.

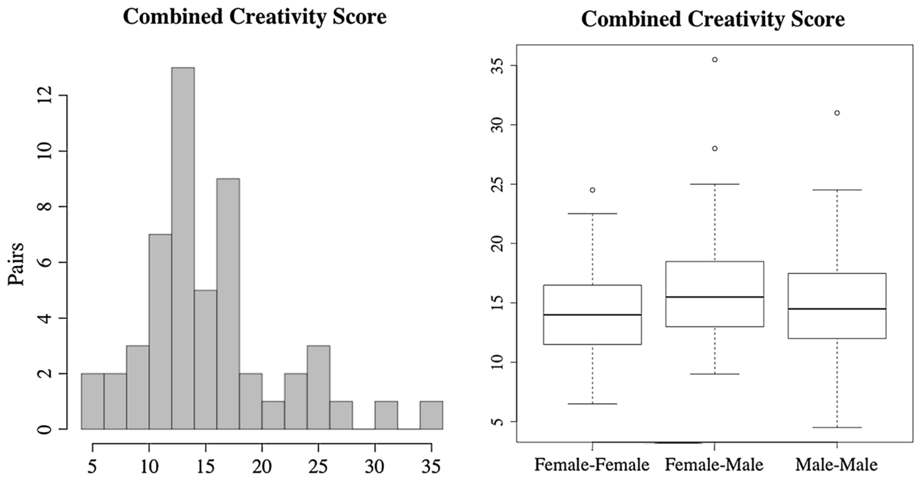


Fig. 3 Histogram shows the distribution of combined creativity scores for pairs, and breakdown of score by the gender composition of pairs

Table 1 Correlations between contributors to creativity

	Quality score	Quantity score	Creativity (sum of quality and quantity)	Word count
Quality score	1	.76	.92	.09
Quantity score	.76	1	.95	.13
Creativity (sum of quality and quantity)	.92	.95	1	.12
Word count	.09	.13	.12	1

Word Count

We wanted to differentiate producing novel, valid ideas from simply speaking, which might correlate with body movement as well. As a result a measure of *word count* was added to take into account a construct that might overlap with our creativity measure. If creative ideas were being spoken aloud, then the physical act of speaking might correlate with the number of ideas spoken. Since fluency, or number of ideas generated, was an important component of our creativity measure, we created a total word count for each interaction ($M = 515.0$, $SD = 170.5$) in order to check whether our creativity measure was confounded with the amount that people talked. Word count was computed by using the Microsoft Word "Word Count" function on the transcripts. Therefore, we also used word count as a second subject of prediction for machine learning analysis.

Given that gender can be a factor in interpersonal interaction, Figs. 3 and 4 also provide visualizations of the outcome measures by the gender composition of the dyads.

Synchrony

The Kinect movement data consisted of the recorded movements of each of the 104 participants. The data were synchronized using the computer time stamp, and in order to make

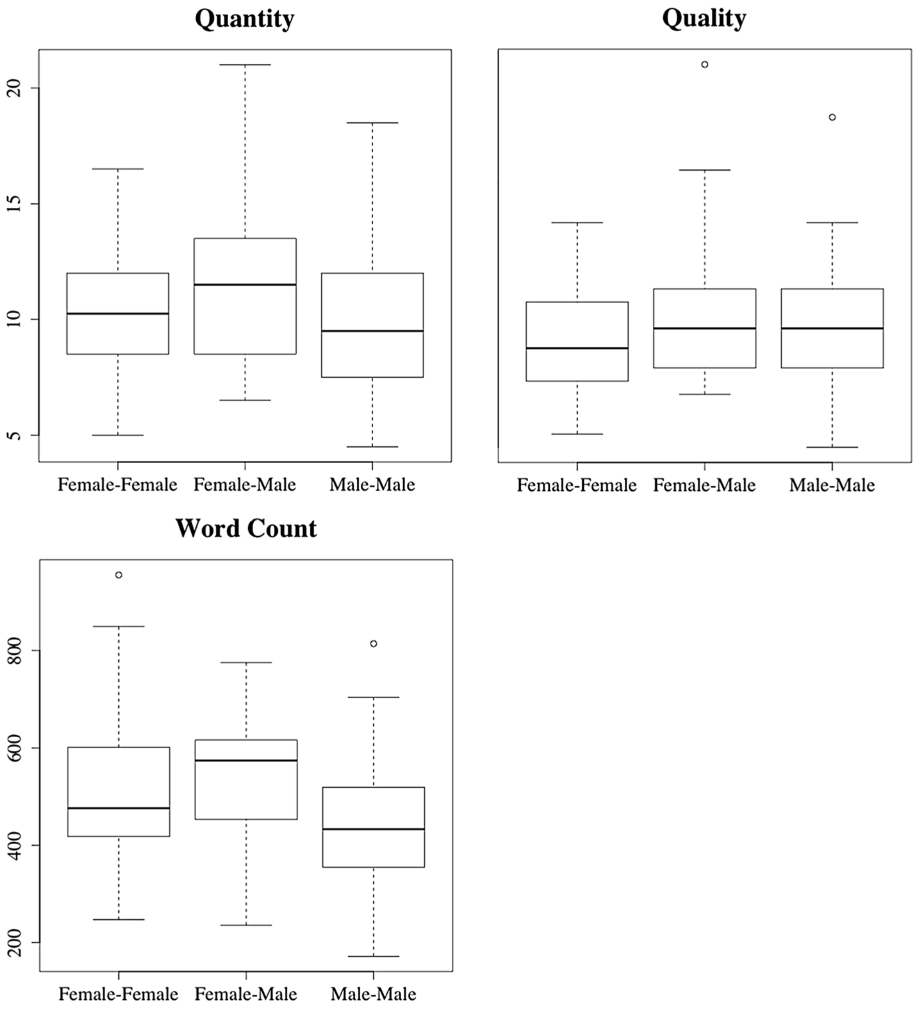


Fig. 4 Summary information for the creativity components and word count differentiated by gender

consistent comparisons of movement, we looked at only the first 3 min (180 s) of each interaction. Since each Kinect recorded an average of 15 frames per second there were thus approximately 2,700 frames per interaction. In each frame, the X Y Z positions of each of the nodes representing a joint of the Kinect skeleton, as well as the overall position of the skeleton, were recorded. The skeleton schematic in Fig. 5 shows the nodes used in our calculations. Four (the foot and hand nodes) were considered to be redundant and were discarded. We also did not use the overall position node, since participants had been instructed to remain on their tape marks and thus their position in the room was constrained. This left us with 16 nodes, or joints, as can be seen in Fig. 5. The output from the Kinect listed whether each node was tracked, inferred (estimated by the algorithm), or missed for every frame. In order to provide a more accurate measure of synchrony, for every frame, we dropped any nodes for which the opposing node was not tracked. Thus, we only used the positions of nodes when the presence—or absence—of synchrony for that node, in that particular frame, could be determined.

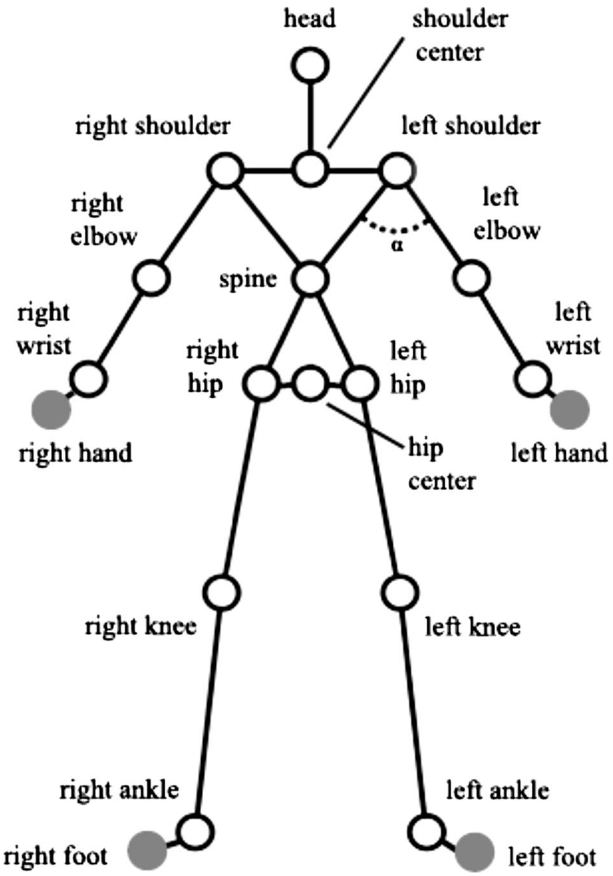
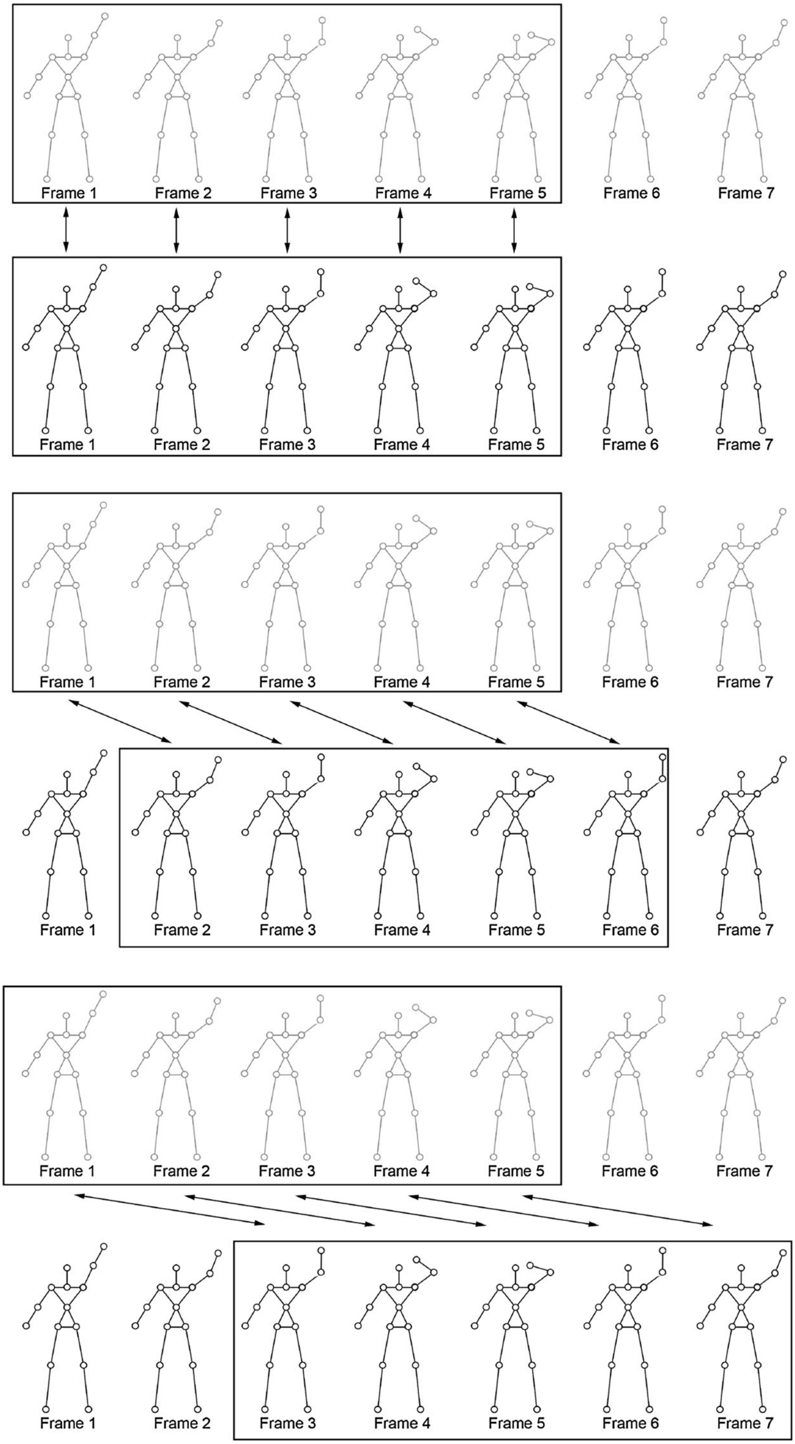


Fig. 5 The modified skeleton derived from Kinect data output in the form of a wireframe. The wireframe consists of 20 nodes with x, y, and z values in physical space. The nodes in *gray* were not used in analysis. Angle α describes an example of an angle used in feature creation; this angle approximates the shoulder and was created from the left elbow, left shoulder and spine nodes

Each skeleton joint could be used to calculate at least one angle, and some could be used in more than one angle, resulting in 18 angles per skeleton. An example angle would be the angle between the line running from the right shoulder node to the right elbow node, and the line running between the right elbow node and the right wrist node, which would correspond approximately to the movements of the right elbow joint.

For each angle that roughly represented a joint, we created an array of each position of the angle, for every second frame, over the entire 180 s time window. As described previously, we were conservative and only used frames for which both corresponding angles were completely tracked, discarding those frames for each angle in which any points in the corresponding angle were not tracked.

For each angle, we computed a *synchrony score* for each pair, which was the correlation between Participant A and Participant B’s movements over a time window of 50 s. This window size was chosen to maximize the amount of data used, while allowing the arrays to be offset in time as described below. Specifically, we calculated the change in each angle from each frame to the next for each participant, reflecting each participant’s movement at


Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7
Frame 1
Frame 2
Frame 3
Frame 4
Frame 5
Frame 6
Frame 7

◀**Fig. 6** The *top row* of this figure shows the movements of participants moving in exact synchrony. Each movement is identical for each frame, and the window size in this example is five frames (as opposed to the 350 used in our actual study). Participant A is shown as the figure in *gray*, and participant B is the figure in *black*. As the window for Participant B is offset, the correlation between the two windows measures an increase in delay

that point in time. Then, again for each angle, we correlated this array of angle movements to the array of movements of the corresponding angle generated by the participant's partner. These 18 correlations represented the synchrony score for that angle, for that pair when their movements were perfectly matched in time.

Following Paxton and Dale's procedure (2013), we then shifted the second participant's window of movements forward in time such that it was offset by one column (two frames, or approximately .13 s) from the second participant's movements. In other words, we examined the relationship between the first participant's movements at one moment in time, and the second participant's movements .13 s later, as expressed by a correlation. We then repeated the shift, and computed another synchrony score, such that we produced 350 correlations for every shift ranging between zero and 50 s. Figure 6 illustrates a small window of this procedure. We did this twice, first shifting the window of movements tracked by the first Kinect forward, and then shifting the window tracked by the second Kinect forward. This was done so that each participant was the "leader" in one time-shift series. We then averaged the resulting two tables of correlations together.

This produced a table of 18 rows of correlations, where each column represented one offset between the two windows of recorded movement. To make our features more meaningful and reduce the need for feature selection when making predictions, we averaged the correlations by body region, grouping them by six body regions: head, torso, right arm, left arm, right leg, and left leg.

In order to illustrate synchrony over time, we present the synchrony scores resulting from 100 consecutive time offsets, for each body region, in Fig. 7. However, whenever we refer to "synchrony scores" as predictive features, we only use the correlation between the movement of each participant's angles from frame to frame at zero offset (i.e., the leftmost points on the X axis in the graphs depicted in Fig. 7). In other words, the synchrony scores as used for prediction are the correlations of movements that occur simultaneously.

Results

First, we sought to assess whether or not we could observe a top-down, intuitively crafted measure of synchrony in the movement data collected from the dyads. We examined how the synchrony score described above decreased as we increased the distance between the two arrays of movements. Next, we assessed whether such a measure of synchrony could predict creativity and word count. We examined the correlations between features, synchrony, and word count (Table 4), and we used machine learning to generate models that could predict the outcome measures using the synchrony scores for each body region. Finally, we examined which features were most important in predicting creativity and word count.

Demonstrating Synchrony

Figure 7 demonstrates that all six features showed a similar pattern over time, such that synchrony decreased as the time lag in the correlation increased. This suggests that

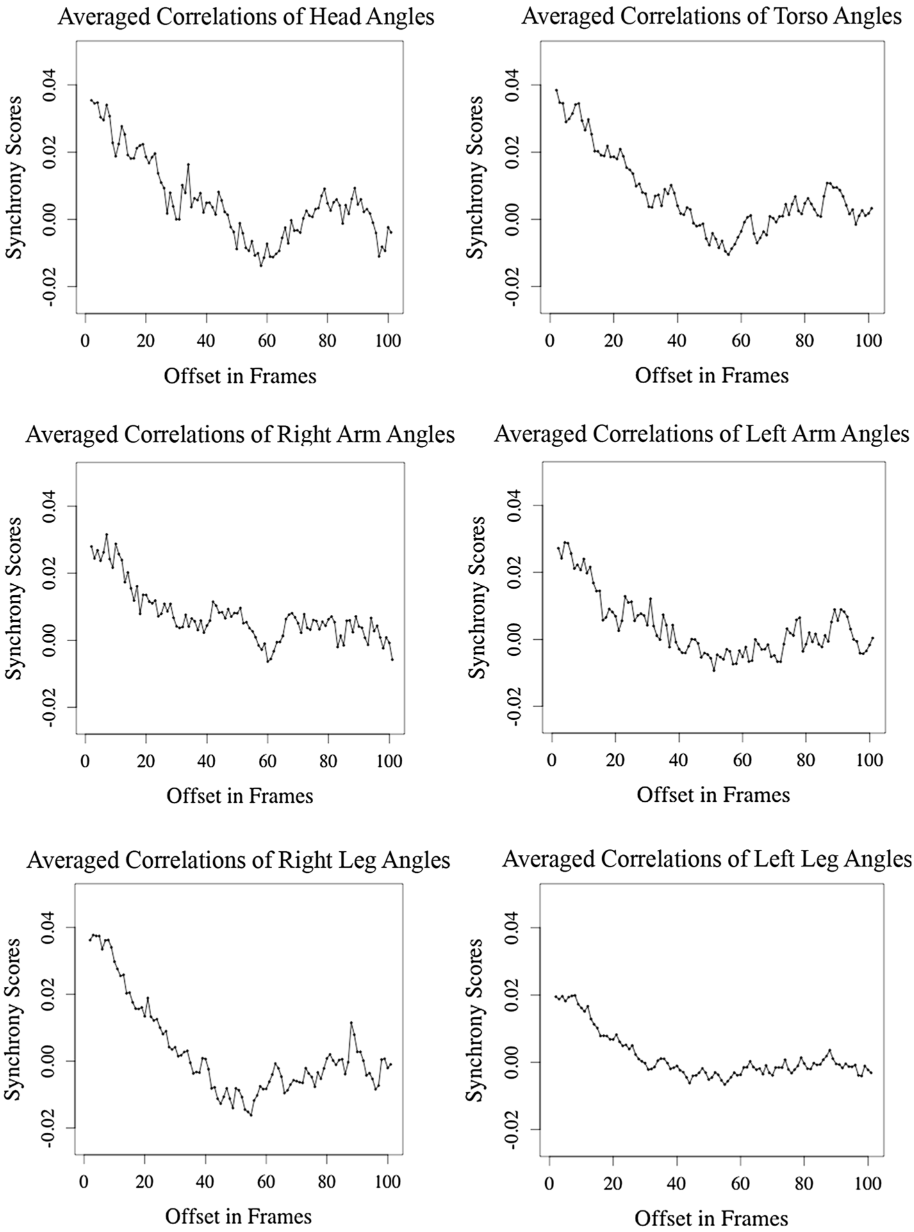


Fig. 7 The X-axis shows the average of correlations of all angles in a given body region, across all pairs, between two 50-s segments of participant movement summary. The Y-axis shows the average of the segments offset in both directions at increasingly large amounts of time, from 0 to 14 s. Synchrony declines from the first offset (of 1 s) becoming nearly flat at around 8 s

movements were indeed in synchrony—as the pairings were further removed in time, the correlations moved toward zero. While the initial correlations are low, typically around .05, the pattern is robust across features and the downward trend is consistent.

Predicting Creativity

We took the first synchrony score—that is, the correlation between body movements that occurred simultaneously for each body region—as input features, providing six features roughly representing major regions of the body- head, torso, right arm, left arm, right leg, and left leg. Table 4 shows the correlations among these features, creativity, and word count.

Following Won et al. (2014) we examined increasingly exclusive datasets of the highest and lowest performing pairs to see if prediction would improve as the cases in the subsets became more and more extreme. In order to compare high and low creativity groups, we wanted to create bins of roughly equal size containing those individuals or pairs whose scores could be categorized as "high" and those whose scores could be categorized as "low." Previous research (Barron 2003) indicates that looking at the differences between high and low scoring groups can provide useful information about the important qualities of an interaction. To be as consistent as possible, we created increasingly exclusive subsets and used the features to predict outcomes on each subset, in order to understand how predictiveness might increase as the divisions between subsets became more extreme.

In order to generate robust results, we used three machine learning algorithms (similar in strategy to Hoque et al. 2012). These were Logistic Regression (LR), Multilayer Perception (MP), and J48 decision tree (J48). Each algorithm used a slightly different approach to finding relationships between the provided features, and used these relationships to predict whether the score of the pair that produced these features would be high or low. For more information on the mechanics of each algorithm, see Witten et al. (2011).

Machine learning uses a bottom-up approach, calculating relationships between the features derived from the data to predict a given outcome. The machine learning algorithms used a train and test procedure to find the most predictive combination of features. For each algorithm in our implementation, we used tenfold cross-validation to estimate the accuracy of the prediction model. Ten-fold cross-validation randomly selects 10 % of the sample to hold out for testing. The rest of the sample is used for training, and the model is tested on the held-out 10 %. This process is repeated ten times (with ten non-overlapping testing datasets) such that the final results combine the results of each of the ten test subsets to get the total score of hits, misses, false positives, and correct rejections. That "confusion matrix" is then combined to produce an accuracy score, in this case simply the percentage of the total number of pairs in each subset whose outcome was correctly predicted (the sum of hits and correct rejections divided by the number of responses).

A typical strategy is to reduce the number of features that are included in the models, which results in increased predictive power and decreased noise, as well as reducing the risk of overfitting. For each model we built, *after* the test cases were removed for each fold, we used correlation-based feature selection to select the features used in the formal modeling, and to reduce the number of features, which were highly correlated with one another. We have reported these results in the text below. However, it should be emphasized that machine learning does not typically provide satisfying explanations about *why* individual features are selected, so causal interpretations offered are necessarily speculative.

We used the synchrony scores of the six body regions as inputs to our models and trained them with the dyad's total creativity score. As can be seen in Fig. 8, accuracies did improve as the subsets became more exclusive, with results averaged across the three classifiers consistently over 70 % in the smaller subsets, and a high score of 86.7 % accuracy. The results for the highest scoring subset are expressed as a confusion matrix in

Table 2. The feature selected most often during feature selection was Head. Plots showing the correlation between this feature and creativity are shown in Fig. 9.

Predicting Word Count

While the main objective of the paper was to predict creativity, we also wanted to address the associated measure of word count, which could also drive the number of creative ideas produced. For our next analysis, we predicted whether participant pairs were classified as high or low word count, using the three algorithms described above. However, as can be seen in Fig. 8, the accuracy of predictions for participants' word count was generally low. The results for the subset of ten pairs are expressed as a confusion matrix in Table 3. It is important to note that null results do not indicate that word count cannot be predicted by body movements. However, this relationship was not very salient in our dataset; we suggest possible explanations for this in the discussion section.

It is notable that all of the correlations between synchrony scores of the six features and word count depicted in Table 4 were negative. In other words, the more that participants spoke overall, the less their movements were in synchrony at zero offset in time. However, since these correlations were not statistically significant, it is best not to draw two many conclusions from these results.

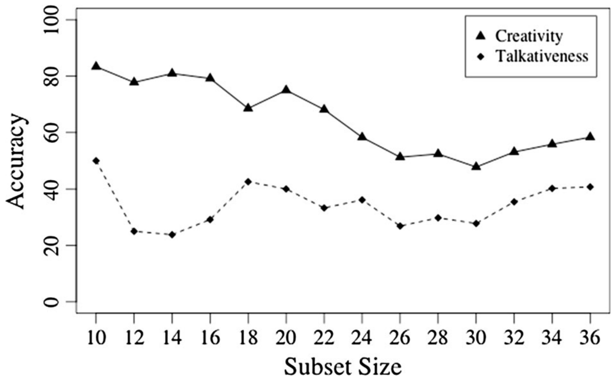


Fig. 8 This plot shows the average accuracy of predictions using three classifiers on increasingly inclusive datasets

Table 2 Predicting creativity from synchrony scores (15 pair subset, 8 high, 7 low)

	Hits	Misses	Correct rejections	False positives	Accuracy (%)
Highest scoring subset					
MP	6	2	7	0	86.7
J48	6	2	6	1	80.0
LR	6	2	5	2	73.3

This confusion matrix was generated using tenfold cross-validation. In order to validate our results we repeated this analysis 10 times, producting an average accuracy of 80 % for J48, 85.4 % for MP, and 74 % for Logistic Regression. For plots showing accuracy scores over increasingly inclusive subsets of high/low creativity pairs, see Fig. 8

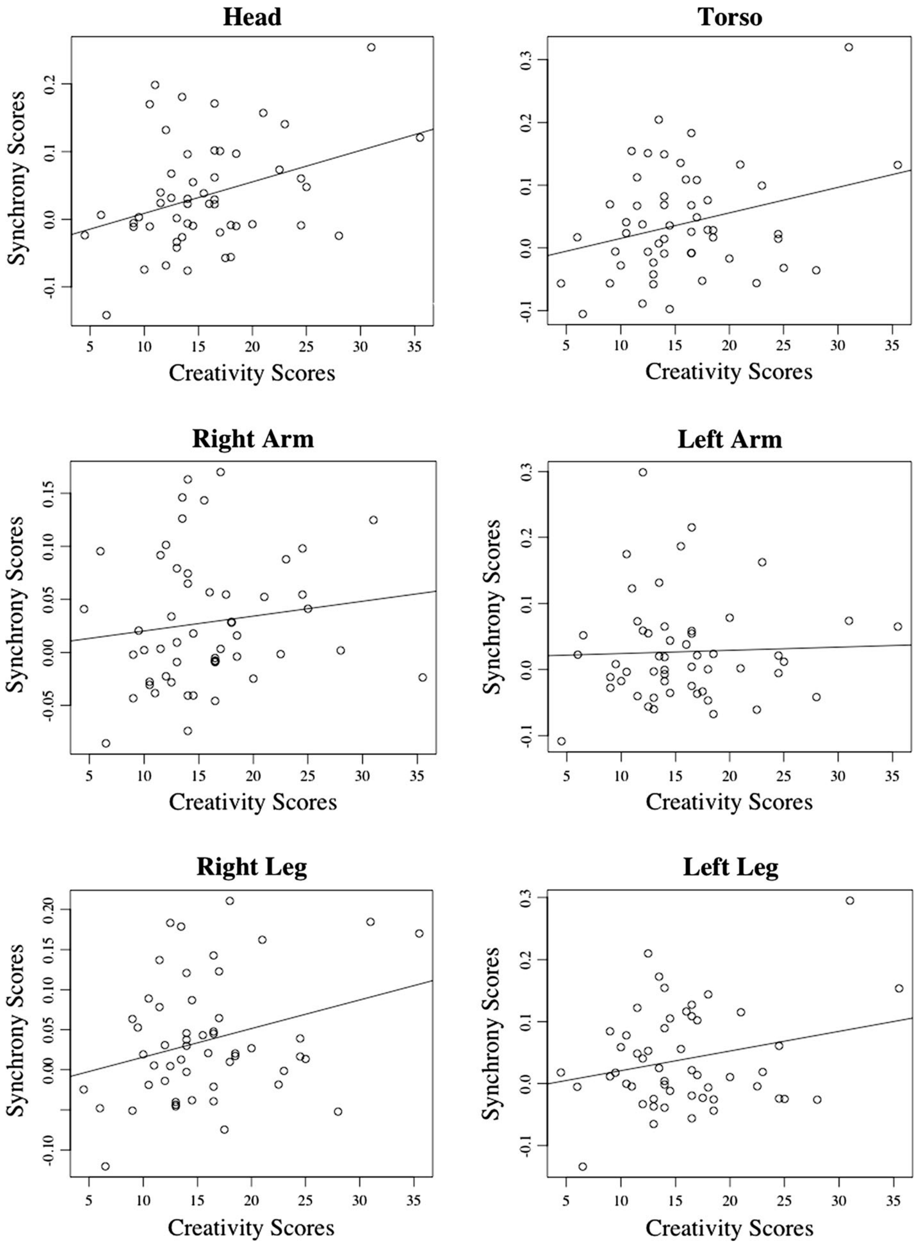


Fig. 9 These plots show the relationship between creativity and all six synchrony scores

Predictive Features

Table 4 demonstrates the synchrony scores for the six body regions correlated with creativity, word count, and each other. The trends shown by the correlations followed the

Table 3 Predicting word count from synchrony scores (10 pair subset, 5 high, 5 low)

	Hits	Misses	Correct rejections	False positives	Accuracy (%)
Highest scoring subset					
MP	1	4	3	2	40.0
J48	3	2	5	0	80.0
LR	0	5	3	2	30.0

These scores were generated using tenfold cross-validation. As with the rest of the subsets for word count, they do not demonstrate predictiveness above chance

trend of feature selection shown with machine learning, which also indicated head as the most predictive feature for creativity.

We thus examined how these features might be distributed among all participants. Plots showing the relationship between each body region's synchrony scores to creativity scores are shown in Fig. 9.

Discussion

In the study described above, we used automatically detected measures of nonverbal behavior to predict success in a creative collaborative task, with a success rate as high as 86.7 %. We used the movement data we collected to investigate the phenomenon of creative collaboration to predict synchrony, and to predict outcomes of a joint interaction both in pair creativity score, and in joint word count of the participants.

We were able to find descriptive trends of synchrony, demonstrating that measures of synchrony declined as the time lag between the nonverbal behaviors of the two dyad members increased. In addition, we were able to predict creativity using measures of synchrony between participants over the course of the interaction, summed by body region. The feature that we found to be most predictive for creativity, head movement, showed a significant linear correlation with score. While we were not able to make consistent predictions for word count, body movements were slightly negatively correlated with word count, which might indicate that participants who talked a lot as a pair did not demonstrate synchronized movement at zero offset. This may be because in conversation, partners will follow each other's movements at a lag, as one participant and then the other takes the lead. While we did not explicitly examine lagged correlations as predictive features in this study, we feel the method of tracking body movements described here may lend itself well to this kind of analysis in the future. It should also be noted that all body movements were highly correlated with one another. Since all features demonstrated synchrony, which features were selected as most predictive may also be a function of our tracking system and method of generating features.

There were several important limitations to our study. First, our method of measuring synchrony was derived from measuring the changes in angles that roughly represent joints. However, these changes in angle do not differentiate between three-dimensional movements. For example, holding one's arms out to the side in a "T" position would create roughly the same angle measurements as holding both arms out directly in front of the body. Thus our features captured the amount of movement over time without referring to specific gestures. Second, our measure of group creativity was imperfect. Although we attempted to find a very general task to measure collaborative cooperation, the fact that the

Table 4 Correlations between significant features and outcome measures

	Creativity	Word count	Head synchrony	Body synchrony	Right arm synchrony	Left arm synchrony	Right leg synchrony	Left leg synchrony
Creativity	1	.12	.36**	.31*	.16	.07	.31*	.28*
Word count	.12	1	−.01	−.13	−.01	−.22	−.09	−.18
Head synchrony	.36**	−.01	1	.72**	.24	.37**	.55**	.52**
Body synchrony	.31*	−.13	.72**	1	.43**	.29*	.78**	.77**
Right arm synchrony	.16	−.01	.24	.43**	1	.20	.31*	.39**
Lett arm synchrony	.07	−.22	.37**	.29*	.20	1	.13	.15
Right leg synchrony	.31*	−.09	.55**	.78**	.31*	.13	1	.82**
Left leg synchrony	.28*	−.18	.52**	.77**	.39**	.15	.82**	1

* Correlations significant at $\alpha < .05$

** Correlations significant at $\alpha < .01$

subject matter was assigned and that most participants were not expert on environmental principles meant that the task was not completely naturalistic. In addition, the usual artificiality of a college educated participant pool between the ages of 18 and 22 must be noted, especially given the fact that group composition, including diversity, may be important in group tasks. Finally, as Kurtzberg and Amabile (2001) point out, group creativity has not been as studied as thoroughly as individual creativity, and other factors yet to be discovered may need to be considered.

The current results reinforce the idea that the quality of an interaction may be predicted in part by body movements, and that the interaction between the movements of both participants is important in predicting interaction outcomes. The fact that these measures can characterize synchrony validates other approaches that look at predictive power of holistic movement (Schmidt et al. 2012; Ramseyer and Tschacher 2011) with the added advantage that it can be readily used to characterize specific body regions.

This paper intends to describe a useful technique in capturing and analyzing movement in interacting pairs, and validate the ability of such systems to make meaningful predictions about the accuracy of such interactions using these recorded movements. However, much work remains to be done in refining these techniques. Some fruitful next steps would be to continue existing work on predicting individual states of mind from gesture, and examine how the combination of individuals' specific gestures might possibly reveal the valence of their interaction. This is also indicated by our result of negative correlations weakly correlating with word count, or "talkativeness." Although these results are not significant, the consistent trend may imply that contingency of gesture is a factor in predicting this outcome. A more nuanced model of back-and-forth synchrony between participants holds the potential to reveal much more about interpersonal interactions.

As Bernieri et al. (1988) point out, "The elements of this [the apparent unification of two behavioral elements into a meaningfully described whole, synchronous event] may be simultaneous, identical, and in phase or alternating, mirrored, and out of phase" (p. 244). In this sense, human coders' ability to recognize synchrony is also a bottom-up process, since the mechanisms by which people recognize synchrony appear to be intuitive and somewhat unconscious. Combining our top-down predictions of what synchrony may entail with bottom-up processing of large data sets may allow us to refine our definitions. The fact that certain body regions were particularly predictive may indicate starting points for future investigation.

Many methods of detecting and evaluating synchrony have been proposed previous to this study (for a review, see Delaherche et al. 2012). The experiment described above reinforces previous research, indicating that automatically detecting nonverbal behavior may allow the prediction of outcomes in dyadic interactions. In addition, it provides additional resources for the definition and measurement of concepts such as synchrony in interpersonal interaction. The increasing interest in social signal processing (Vinciarelli et al. 2008) and the concept of production and perception of nonverbal behavior containing cross-cultural elements (Zebrowitz and Montepare 2006) suggest that methods to automatically detect and measure nonverbal behaviors based on body movements will continue to be an important area of study.

Acknowledgments The work presented herein was funded in part by Konica Minolta as part of a Stanford Media-X grant, and we thank them for the valuable insights provided by their visiting researchers, in particular Dr. Haisong Gu. In addition, it was funded in part by grant 108084-5031715-4 from the National Science Foundation. The authors also thank lab manager Cody Karutz for coordinating the administrative aspects of running this study, and Jimmy Lee, Pamela Martinez, Evan Shieh, Alex Zamoshchin, Angel Olvera, Christine Tataru, Mark Diaz and Mark Peng for their help in coding and data analysis, and Dr. Laura

Aymerich-Franch, Ketaki Shriram, Michelle Friend, Brian Perone, and Jakki Bailey for helpful comments on an earlier draft of this paper.

References

Baas, M., De Dreu, C. K., & Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin, 134*(6), 779.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*(3), 307–359.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher student interactions. *Journal of Nonverbal Behavior, 12*, 120–138.

Bernieri, F. J., Davis, J. M., Rosenthal, R., & Knee, C. R. (1994). Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect. *Personality and Social Psychology Bulletin, 20*(3), 303–311.

Bernieri, F. J., Reznick, J. S., & Rosenthal, R. (1988). Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. *Journal of Personality and Social Psychology, 54*(2), 243.

Castellano, G., Villalba, S. D., & Camurri, A. (2007). Recognising human emotions from body movement and gesture dynamics. *Affective Computing and Intelligent Interaction, 4738*, 71–82.

Condon, W. S., & Ogston, W. D. (1966). Sound film analysis of normal and pathological behavior patterns. *Journal of Nervous Mental Disorders, 143*, 338–347.

Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., & Cohen, D. (2012). Interpersonal synchrony: A survey of evaluation methods across disciplines. *Affective Computing, IEEE Transactions on, 3*(3), 349–365.

Drolet, A. L., & Morris, M. W. (2000). Rapport in conflict resolution: Accounting for how face-to-face contact fosters mutual cooperation in mixed-motive conflicts. *Journal of Experimental Social Psychology, 36*(1), 26–50.

Goncalo, J. A., & Staw, B. M. (2006). Individualism–collectivism and group creativity. *Organizational Behavior and Human Decision Processes, 100*(1), 96–109.

Grahe, J. E., & Bernieri, F. J. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior, 23*(4), 253–269.

Guilford, J. P. (1957). Creative abilities in the arts. *Psychological Review, 64*(2), 110–118.

Harrigan, J. A., Oxman, T. E., & Rosenthal, R. (1985). Rapport expressed through nonverbal behavior. *Journal of Nonverbal Behavior, 9*(2), 95–110.

Hoque, M. E., McDuff, D. J., & Picard, R. W. (2012). Exploring temporal patterns in classifying frustrated and delighted smiles. *Journal of IEEE Transactions on Affective Computing, 99*, 1–13.

Huang, L., Morency, L. P., & Gratch, J. (2011). Virtual rapport 2.0. In *Intelligent virtual agents* (pp. 68–79). Berlin: Springer.

Jabon, M. E., Ahn, S. J., & Bailenson, J. N. (2011a). Automatically analyzing facial-feature movements to identify human errors. *IEEE Journal of Intelligent Systems, 26*(2), 54–63.

Jabon, M. E., Bailenson, J. N., Pontikakis, E. D., Takayama, L., & Nass, C. (2011b). Facial expression analysis for predicting unsafe driving behavior. *IEEE Pervasive Computing, 10*(4), 84–95.

Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., & Driessen, P. F. (2005). Gesture-based affective computing on motion capture data. In *Affective Computing and Intelligent Interaction* (pp. 1–7). Berlin, Heidelberg: Springer.

Kendon, A. (1970). Movement coordination in social interaction: Some examples described. *Acta Psychologica, 32*, 100–125.

Kleinsmith, A., & Bianchi-Berthouze, N. (2007). Recognizing affective dimensions from body posture. In *Affective computing and intelligent interaction* (pp. 48–58). Berlin: Springer.

Kurtzberg, T. R., & Amabile, T. M. (2001). From Guilford to creative synergy: Opening the black box of team-level creativity. *Creativity Research Journal, 13*(3–4), 285–294.

La France, M., & Broadbent, M. (1976). Group rapport: Posture sharing as a nonverbal indicator. *Group and Organization Studies, 1*, 328–333.

Lumsden, J., Miles, L. K., & Macrae, C. N. (2012). Perceptions of synchrony: Different strokes for different folks? *Perception, 41*(12), 1529.

Martin, C. C., Burkert, D. C., Choi, K. R., Wieczorek, N. B., McGregor, P. M., Herrmann, R. A., et al. (2012, April). A real-time ergonomic monitoring system using the Microsoft Kinect. In *IEEE Systems and Information Design Symposium* (SIEDS) (pp. 50–55).

McLeod, P. L., Lobel, S. A., & Cox, T. H. (1996). Ethnic diversity and creativity in small groups. *Small Group Research, 27*(2), 248–264.

Meservy, T. O., Jensen, M. L., Kruse, J., Burgoon, J. K., & Jay, F. (2005). Detecting deception through automatic, unobtrusive analysis of nonverbal behavior. *IEEE Intelligent Systems, 20*(5), 36–43.

Oppezzo, M., & Schwartz, D. L. (2014). Give your ideas some legs: The positive effect of walking on creative thinking. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. doi:10.1037/a0036577.

Paxton, A., & Dale, R. (2013). Frame-differencing methods for measuring bodily synchrony in conversation. *Behavior Research Methods, 45*(2), 329–343.

Pentland, A. S. (2010). *Honest signals*. Cambridge: MIT press.

Ramseyer, F., & Tschacher, W. (2011). Nonverbal synchrony in psychotherapy: Coordinated body-movement reflects relationship quality and outcome. *Journal of Consulting and Clinical Psychology, 79*(3), 284–295.

Schmidt, R. C., Morr, S., Fitzpatrick, P., & Richardson, M. J. (2012). Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior, 36*(4), 263–279.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). *Human activity detection from RGBD images*. AAAI 2011 workshop: Plan, activity, and intent recognition.

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry, 1*(4), 285–293.

Torrance, E. P. (1970). Influence of dyadic interaction on creative functioning. *Psychological Reports, 26*(2), 391–394.

Vinciarelli, A., Pantic, M., Bourlard, H., & Pentland, A. (2008). Social signal processing: State-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on multimedia* (pp. 1061–1070). New York: ACM.

Witten, I., Eibe, F., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques*. Burlington MA: Morgan Kaufman.

Won, A. S., Bailenson, J. N., & Janssen, J. H. (2014). *Automatic detection of nonverbal behavior predicts learning in dyadic interactions*. Manuscript submitted for publication.

Microsoft Corp. Redmond WA. Kinect for Xbox 360.

Zebrowitz, L. A., & Montepare, J. M. (2006). The ecological approach to person perception: Evolutionary roots and contemporary offshoots. In M. Schaller, J. A. Simpson, & D. T. Kenrick (Eds.) *Evolution and social psychology*. First Edition (pp. 81–113), New York: Psychology Press.