# Critical Features of Joint Actions that Signal Human Interaction

**Tianmin Shu** [1] **Steven Thurman** [1] **Dawn Chen** **Song-Chun Zhu** **Hongjing Lu**
{tianmin.shu, sthurman}@ucla.edu sdawnchen@gmail.com sczhu@stat.ucla.edu hongjing@ucla.edu
Departments of Psychology and Statistics, University of California, Los Angeles

## Abstract

We examined the visual perception of joint actions, in which two individuals coordinate their body movements in space and time to achieve a joint goal. Animations of interacting action pairs (partners in human interactions) and non-interacting action pairs (individual actors sampled from different interaction sequences) were shown in the experiment. Participants were asked to rate how likely the two actors were interacting. The rating data were then analyzed using multidimensional scaling to recover a two-dimensional psychological space for representing joint actions. A descriptive model based on ordinal logit regression with a sparseness constraint was developed to account for human judgments by identifying critical features that signal joint actions. We found that identification of joint actions could be accomplished by assessing inter-actor correlations between motion features derived from body movements of individual actions. These critical features may enable rapid detection of meaningful inter-personal interactions in complex scenes.

**Keywords:** joint action; feature selection; human interaction

## Introduction

Humans actions provide critical information for understanding other people's intentions and reacting accordingly. Beyond recognition of individual actions, the ability to engage in *joint* action (i.e., when two or more individuals coordinate their actions in space and time) paves the way for other social interactions (Sebanz, Bekkering, & Knoblich, 2006; Thurman & Lu, 2014). In everyday life, we constantly coordinate our own actions with those of others to achieve a joint outcome involving a change in the environment (e.g., lifting a box together and moving it to a different location), or to achieve a social goal through human interaction (e.g., walking towards one another and giving a "high-five" as a greeting). Carpenter (2009) provided evidence that the ability to participate in joint action is already fairly developed after the first year of life.

However, it is by no means a trivial task to identify whether two persons are interacting in a meaningful manner solely from visual input. There are many circumstances in which individuals incidentally intersect the orbit of other people's body movements, creating a scene that might be confusable with potential interactivity coordinated by the two people. Hence, it is important to examine what specific information in a visual input signals engagement in joint action, and to determine the conditions under which joint action emerges from body movements of individuals.

In the literature, both reasoning-based and feature-based mechanisms have been proposed to support the recognition and planning of actions when interacting with the environment or other people. In particular, there have been several recent efforts to interpret an agent's behavior via the intentional stance required to understand joint actions. For example, Baker et al. (2009) developed a computational model for reasoning about intentions within a sprite world inspired by the seminal work of Heider and Simmel (1944), in which simple shapes (e.g., red circles or blue triangles) move around in a constrained environment. Although these studies illustrate the potential fruitfulness of applying high-level constraints to reason about the goal underlying observed actions, the investigations have been limited to simplified environments and movements of rigid objects. Relatively few psychological studies have used whole-body movements of humans as the visual input to examine the mechanisms underpinning joint actions.

On the other hand, Marsh et al. (2009) proposed that social interaction through joint actions can be understood as emerging from dynamical principles across individuals, rather than relying on explicit reasoning on intention, at least in some situations. For example, Richardson et al. (2007) showed that connections between humans can arise through synchronization of action patterns (e.g., rocking together when sitting on rocking chairs side-by-side), rather than mental state attribution. This line of work suggests the possibility that bottom-up feature-based mechanisms may play an important role in signaling joint actions. However, the related studies have been limited to uninstructed coordination of simple incidental rhythmic movements, which only cover a small range of possible human actions, and did not aim to probe the important features underpinning meaningful interactions between actors. In the literature of action recognition, there is rich evidence showing that humans are sensitive to some signature movements to enable efficient action detection (Casile & Giese, 2005; Troje & Westhoff, 2006; van Boxtel & Lu, 2015, 2012). Hence, it is conceivable that some critical features of coordinated movements between actors can facilitate the perception of joint actions.

The present study investigated people's mental representations of joint actions and how well features measuring different types of movement coordination between individual actors can capture human judgments of interaction. We collected human ratings of the degree to which two actors seem to be interacting with each other for different video clips. These data were then used to derive a psychological space representing joint actions between two individuals. Computational models were developed to pinpoint critical features of coordinated movements used in constructing a representation space for joint actions by fitting the models to human interactivity ratings.

[1] These two authors contributed equally.

## The Experiment

The experiment was designed to investigate perception of joint actions between two human actors using naturalistic stimuli. Two human actors were recorded engaging in various forms of social interaction, such as shaking hands, passing a water bottle, and salsa dancing. We then generated the full set of stimuli by pairing each actor not only with their true interaction partner, but also with each of the other actors involved in different recorded joint actions (e.g., a salsa dancer on the left with a person shaking hands on the right). We sought to determine the extent to which subjects would recognize true interpersonal interactions through joint action, and also whether the stimulus properties of non-interacting stimulus pairs would result in coincidental visual cues that signal attribution of social interaction in joint action.

## Method

### Stimuli

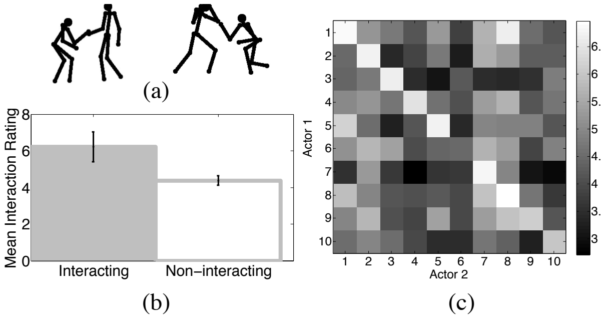
Action stimuli were generated from the CMU Mocap database (`http://mocap.cs.cmu.edu`) and processed by the Biological Motion Toolbox (van Boxtel & Lu, 2013). We selected ten paired interactions in which two human agents were engaged in various social interactions (see Table 1). A total of 100 stimuli with paired actions were presented in the study, including 10 truly interacting with coupled partners and 90 not interacting. The point-light stimuli were rendered as stick figures with lines connecting the joints, and videos lasted 3.67 seconds (110 frames presented at 30 fps).

### Procedure

Participants first viewed two videos, one of a single stick figure walking and another running, and were asked to write a description of what each person was doing. This was to ensure that they understood the format of the videos and were sufficiently competent in English to complete the remainder of the experiment. Participants then viewed 25 videos, 5 of which were chosen from the 10 interacting pairs and 20 of which were the non-interacting pairs formed from the remaining 5 actors. For each video, participants were asked to "rate the degree to which the actors appear to be interacting" on a scale from 1 (Definitely NOT) to 7 (Definitely).

In addition to the 25 videos of interest, participants viewed two videos that were presented on randomly selected trials among the other videos in order to check whether they were paying attention to the task. These were videos of a single actor walking or running. Participants were asked to use a slider (as in the rating questions) to choose the action depicted in each video. Participants who failed either of these attention checks were excluded from the data analysis.

After the video rating task, participants completed the Autism Quotient (AQ) questionnaire (Baron-Cohen et al., 2001), where an attention check was also included.



(a)

(b)

(c)

Figure 1: (a) Frames of two confusing stimuli: 1) arm wrestling and shaking hands, 2) threatening and arm wrestling. (b) Mean interactivity ratings for truly interacting and non-interacting stimuli, represented by collapsing across action types. (c) The matrix of mean interactivity ratings for all pairs, highlighting the variable ratings given to different non-interacting stimuli.

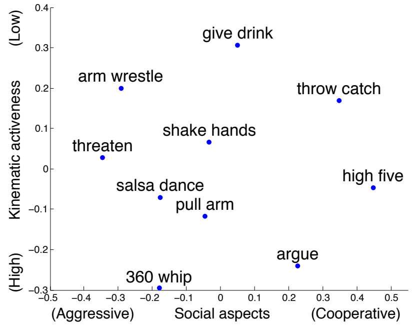


Figure 2: MDS solution derived from human interactivity ratings. The vertical dimension captures activeness of body movements involved in joint action, whereas the horizontal dimension associates with social aspects of joint actions with the exception of "argue" sequence, in which the social content in the point-light display is ambiguous (as this joint action can be interpreted either as an argument or as a normal conversation between two people).

### Results

195 participants were recruited and paid $0.80 for participation in the 5-10 minute experiment on Amazon's Mechanical Turk (Mturk). We excluded 26 participants because they failed one or more of the attention checks during the experiment. Furthermore, we included two additional exclusion criteria to remove participants who did not pay attention to the task. The first measure was the standard deviation of each participant's interactivity ratings across all the testing trials. We excluded 5 participants based on the standard deviation of their ratings being less than 2 z-scores below the mean standard deviation for all participants. Finally, 4 participants were excluded for mistakenly providing higher average interactivity ratings for the non-interacting pairs than for the interacting pairs. Data from the remaining 160 participants were analyzed.

As shown in Figure 1b, human participants yielded significantly higher mean ratings for interacting pairs (mean $= 6.21 \pm 0.82$) than for non-interacting pairs (mean $= 4.38 \pm 0.27$), demonstrating that Mturk participants were sensitive to true joint actions supporting meaningful human interactions (see Figure 1b). Figure 1c showed the mean interactivity rating for each of the 100 pairs of actors. These results can be visualized as an interactivity-rating matrix in which the diagonal elements represent mean ratings for the true interaction stimuli, and the off-diagonals represent mean ratings for non-interacting stimuli. There was substantial variability in interactivity ratings for the set of non-interacting stimuli, with some pairs (e.g., two examples in Figure 1a) yielding high ratings, and other pairs (e.g., giving a high-five and arm wrestling) yielding relatively lower ratings.

We analyzed the matrix of mean interactivity ratings using multidimensional scaling (MDS). MDS is usually applied to similarity data, and the rated interactivities for different pairs of actions can be viewed as analogous to similarities. The MDS algorithm requires that the input matrix is symmetric and that its diagonal elements are all ones (i.e., items are maximally similar to themselves). We therefore divided every element in the rating matrix by the maximum rating along the diagonal, then set all diagonal elements to 1, and finally took the average of the upper and lower triangles of the matrix to form a symmetric input matrix.

Figure 2 shows the results of applying multidimensional scaling to the rating matrix. Because we used interactivity rather than similarity ratings, actions that are closer in this space are more likely to be rated as interactive when paired together. Note that "shake hands", "salsa dance", and "pull arm", which are close to one another, are all actions with sustained touch. In addition, the two dimensions appear to reveal distinct psychological variables. The vertical dimension corresponds to the levels of activity involved, ranging from highly active actions (e.g., "play 360 whip", in which two people hold hands and jump while rotating 360 degrees) to less active actions (e.g., "give drink", in which one standing person passes a can of soft drink to another person standing next to him). The horizontal dimension appears to be associated with social aspects of joint actions, ranging from aggressive, threatening actions (e.g., one person holds an invisible object and appears to threaten another person), to more friendly and cooperative joint actions (e.g., two people walking towards each other to give a high five). One exception was the "argue" joint action, in which two people presumably in argument raise their arms, but never touch each other. When this action sequence is converted into a point-light display, the social content becomes ambiguous (as this joint action can be interpreted either as an argument or as a normal conversation between two people). This is confirmed by the result that the interactivity rating for the "argue" joint action was also the lowest among all the true interactions.

In addition, we found a statistically significant correlation between AQ score and discrimination score, $r = -0.17$, $p = .024$. This relationship suggests that participants with more autistic traits were less able to distinguish between interacting and non-interacting stimuli than participants with fewer autistic traits. This finding is consistent with other evidence of impaired social cognition in autism.

# The Model

In order to determine what visual features play important roles in guiding human judgments regarding meaningful interactions between individuals, we developed a descriptive model based on ordinal logit regression, coupled with a sparseness constraint to encourage selection of a relatively small number of discriminative features. We first describe the model used to predict interactivity judgments for experimental stimuli, and its selection of critical features. We will then provide details on the computation of five different types of features from body movements.

## Rank Prediction and Feature Selection

The model predicts a rank order of interactivity judgments for each pair of actions shown in the stimulus based on the 2D coordinates of joints involved in individual actions. To train the model, human interactivity ratings were converted to rank order. We clustered average human ratings from the entire experiment into four rank levels, with equal numbers of cases assigned to each level. After training, the model predicts the rank order of the interactivity judgment $y$ for a given unseen pair of actions, $y = 1, \cdots, 4$, based on the visual features $\boldsymbol{x}$ derived from the input video. Based on the logit link, we can write the conditional probability for a rank $j = 1, \cdots, 4$ as

$$\pi_j(\boldsymbol{x}) = p(y = j \mid y \leq j, \boldsymbol{x}) \; = \; \frac{\exp(\psi_j + \boldsymbol{x}^\top \beta)}{1 + \exp(\psi_j + \boldsymbol{x}^\top \beta)}, \qquad (1)$$

where $\{\psi_j\}_{j=1,\cdots,4}$, thresholds for the rank $j$, and $\beta$, the coefficients for the features are the parameters to be learned. We use a dummy variable $y_i^j \in \{0,1\}$ to indicate whether the rank of the $i$-th instance in the data is $j$. Then for each instance, we define a vector $\boldsymbol{y}_i = (y_i^1, \cdots, y_i^4)^\top$. Let $Y = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n)^\top$ be the reconstructed responses of the training instances. For $n$ training instances, the log-likelihood is defined as

$$\begin{aligned} & \ell(\beta, \{\psi_j\}_{j=1,\cdots,4}|X,Y) \\ = \; & \sum_{i=1}^{n} \sum_{j=2}^{4} y_i^j \log \pi_j(\boldsymbol{x}_i) + (1 - \sum_{j'=j}^{4} y_i^{j'}) \log(1 - \pi_j(\boldsymbol{x}_i)). \end{aligned} \qquad (2)$$

Since a large number of features are included in the analysis, the model selects the critical features according to a sparseness constraint that includes an $\ell_1$ norm penalty for the coefficients $\beta$. Hence, for $n$ instances with $p$-dimensional feature vectors $X$ and labels $Y$, we maximize

$$\ell(\beta, \{\psi_j\}_{j=1,\cdots,4}|X,Y) - \lambda \sum_k |\beta_k|. \qquad (3)$$

Finally, given learned model parameters (i.e., $\beta$ and $\psi$) and the features of a new test stimuli $\boldsymbol{x}$, the estimated rank $\hat{y}$ of interactiveness for the test stimuli can be obtained based on the following probability for each $j = 1, \cdots, 4$:

$$p(y = j|\boldsymbol{x}) = \begin{cases} \pi_4(\boldsymbol{x}) & \text{if } j = 4 \\ \prod_{j'=2}^{4}(1 - \pi_{j'}(\boldsymbol{x})) & \text{if } j = 1 \\ \pi_j(\boldsymbol{x}) \prod_{j'=j+1}^{4}(1 - \pi_{j'}(\boldsymbol{x})) & \text{otherwise} \end{cases} \qquad (4)$$

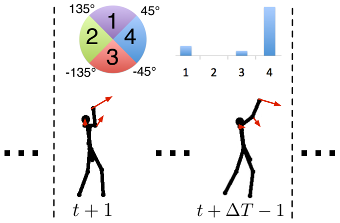


Figure 3: Illustration of motion entropy for the right limb. In a time window of $\Delta T$, we compute the moving directions of the three joints on the right limb at each frame, and compute the histogram of moving directions, which is re-weighted by the magnitude of the velocities. The entropy of the histogram is the motion entropy of the right limb within the time window.

### Features of Coordinated Body Movements

A total of 172 features were provided for the model to select. The features are derived from the coordination of limb movements between two actors. We grouped the features into five types according to their functional roles.

**Type I: Touching**. When viewing a pair of actions, one actor touching the other actor could signal potential joint action with interactive activity (e.g., *Shake Hands*, *Give Drink*, etc.). The feature of touching can be quantified using spatial distance of two joints, each from one actor in the stimulus. The spatial distance of joints can be denoted as $d^t_{ij}$, where $i$ and $j$ denote index of body joints from the two individuals respectively, and $t$ indicates the frame number. An auxiliary variable $T_{ij} = \sum_t \mathbb{1}(d^t_{ij} < D)$ is introduced to count the number of frames when a pair of joints between the two actors are closer than a threshold, i.e., $D = 8$ pixels. The two joints are considered to be "touching" each other if $T_{ij}$ is longer than $\tau = 30$ frames (i.e., 1 second), which can be denoted as a dummy variable $s_{ij} = \mathbb{1}(T_{ij} > 30)$. Then we can obtain the total number of "touching" pairs of joints by $S = \sum_{ij} s_{ij}$. To eliminate accidental spatial overlap of joints when viewing the action pairs from a particular viewpoint, the two actors are defined to be touching at certain point if and only if one or two pairs of joints between them are considered to be touching, i.e., $1 \leq S \leq 2$. The touching variable is set as 1 if the two actors' joints satisfy this criterion.

**Type II: Passing**. In some joint actions supporting human interactions, two actors can pass each other, producing spatial overlap of the two actors. The second type of features aim to capture spatial relations between two actors in the stimulus. Similar to the touching feature, we can also count $S$ using the same method as defined in the previous paragraph with thresholds $D = 13$ and $\tau = 30$. The binary passing variable is coded as 1 indicating the presence of this feature if $S \geq 3$.

**Type III: Temporal correlation of body movements**. The third type of features aims to capture the variability of limb movements over time. Here we define four limbs for an actor (left arm, right arm, left leg and right leg), with each limb including 3 joints. For example, if a person only moves the upper body in joint action (e.g., passing an object while sitting on a chair), there is no variability of leg movements over time, but the arm movements change when actions unfold over time. To quantify this, for each limb we compute entropy based on the histogram of moving directions of joints in two-dimensional space. We first equally quantize joint motion directions into 4 bins as shown in Figure 3, and then count the frequency of moving directions of each joint on that limb at each frame in the given temporal window $\Delta T = 6$. To ensure sensitivity to the influence of joint movement speed, the frequency counts are weighed by velocity magnitude of joint movements. If the joints move in a similar direction over time or remain static, the entropy measure during this period is low. In contrast, the entropy is high when the joints move in varied directions and speeds over time. Because actions progress over time, we obtain a time series of motion entropy for each limb of individual actors in stimuli. To capture the coordination of body movements in joint actions, we compute the correlation of motion entropy sequences between the two actions for each limb pair, resulting in 16 features (i.e., all possible combination of four limbs in each actor). In addition, we include another feature to capture the temporal correlation of whole-body motion entropy sequences for the two actors, where entropy is calculated by pooling motion from all joints. In total, set III has 17 features.

**Type IV: Correlation of motion trajectories**. Two actors engaged in joint action often perform the same movements at the same time, e.g., raising arms together to give a high-five. To capture this kind of limb movement coordination, we calculated the inter-actor correlation of the motion trajectories for the centers of mass of the four limbs and for the center of mass of the entire body, yielding a total of 17 features.

**Type V: Motion correlation with temporal shifts**. A characteristic of joint actions is that one person acts in response to the other person's actions. In some situations, limbs move in a synchronized manner (e.g., each actor lifting arms in synchrony to give a high-five greeting), which is capture by feature set IV. In other situations, however, one person moves their limbs first to initiate an interaction (e.g., a person passing an object to the other person). To capture these temporal relations in coordinated movements, we introduced a range of relative temporal shifts (-1.6s, -0.8s, 0.8s, 1.6s) of motion trajectories between two actors to calculate the correlation indices defined in sets III and IV, yielding 136 features in set V.

Finally, all features were standardized to fix the mean at 0 and the variance at 1 across all training stimuli. Any feature values within 1.2 standard deviations of the mean were set to the mean value in order to remove insignificant feature values and to facilitate the feature selection process.

## Model Results

### Rank Predictions for Unseen Interactions

To evaluate whether the model can generalize its learned features to new joint actions, we split the 100 stimuli into two sets for each of the 10 joint actions: a training set with 81 pairs formed from the other 9 actions, and a testing set with the remaining 19 pairs, one of which is the truly interacting

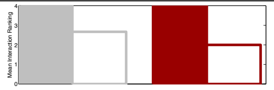
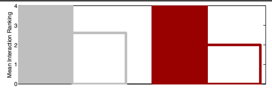
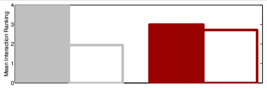
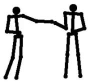
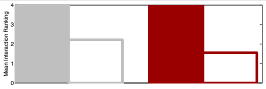
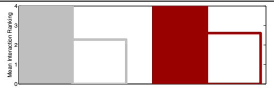
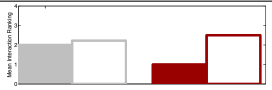
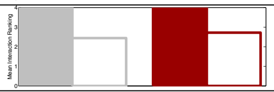
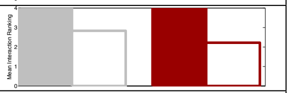
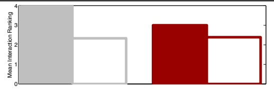
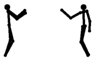
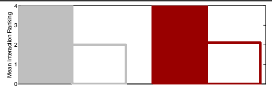
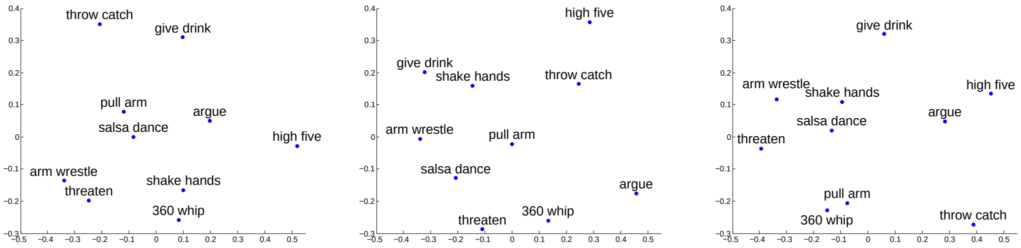
| ID | Name | Illustrative Frame | Mean Interaction Ranks | ID | Name | Illustrative Frame | Mean Interaction Ranks |
|---|---|---|---|---|---|---|---|
| 1 | Shake Hands | |  | 2 | Pull by Arm | |  |
| 3 | High Five | |  | 4 | Give Drink |  |  |
| 5 | Arm Wrestle | |  | 6 | Argue | |  |
| 7 | Play 360 Whip | |  | 8 | Salsa Dance | |  |
| 9 | Threaten | |  | 10 | Play Catch |  |  |

Table 1: Comparison between mean human interactivity ranks and model predictions for each type of interaction. An illustrative frame from each interaction video is also shown. The gray and red bars are the average ranks based on human ratings and model predictions respectively. Solid bars denote the ranks for interacting pairs and blank bars denote the average ranks for the non-interacting pairs in that testing set.



(a) The first three feature sets.

(b) The first four feature sets.

(c) All features.

Figure 4: MDS results of fitted ranks from models trained with different feature sets.

pair performing the joint action of interest and 18 of which are non-interacting pairs formed from the partners in the joint action of interest coupled with actors from different joint actions. The final model for each of the 10 joint actions is fitted to select 22 critical features, which yielded the highest correlation between predicted ranks and human ranks.

Table 1 depicts the model's prediction when tested on each interactive pair, in comparison to rank orders from human ratings. The overall correlation between model predictions and human judgments is strong, with $r = .85, p < 10^{-5}$. The average root-mean-square error (*RMSE*) of rank order was 0.47. The best fitted joint actions are *Play Catch*, *Play 360 Whip*, and *Arm Wrestle*, which yielded low *RMSE* (0.08, 0.20, and 0.24 respectively), and the worst fitted actions were *High Five*, *Argue*, and *Threaten* with high *RMSE* (0.90, 0.73, and 0.71 respectively). These differences in *RMSE* across various joint actions suggest that the descriptive model using features of coordinated movements is able to predict the rank order of joint actions which can be defined with strong visual cues, while the model shows its limitation to other joint actions which are rich in social content (such as threaten) but less informative in terms of coordinated movements.

The five features commonly selected across different training sets are 1) touching, 2) correlation between the motion entropies of the right leg of actor 1 and the left arm of actor 2 with an 0.8s temporal shift, 3) correlation between the trajectories of the left leg of actor 1 and the left leg of actor 2 with a -1.6s temporal shift, 4) correlation between the whole-body motion entropies of the two actors, and 5) correlation between the trajectories of the right arm of actor 1 and the left arm of actor 2, which were all chosen on at least 8 test runs with positive coefficients. In particular, touching was selected on all test runs, showing its generalization to a large range of joint actions. These results regarding the features commonly selected by the model are robust, as feature rankings do not change significantly across different sets of model parameters.

## MDS from fitted ranks

We trained the model with all 100 action pairs and let the model predict the interactiveness of any pair of actions so that we can derive the model-fitted rating matrix, which was used to recover the psychological space of joint actions by the multidimensional scaling algorithm. The joint action space derived from the model-simulated matrix makes it possible to examine how different types of features impact similarity between different joint actions. We fit the rating matrix using the first three feature sets, first four feature sets and all features in the five sets, respectively. The resulting MDS results are shown in Figure 4. The comparison of model-derived joint action space with human results shows that set V features play the most important role in mimicking human judgments, as one dimension corresponds to the amount of visual motion information in observed action pairs, and the other dimension is associated with social aspects of joint action. Given that the model is solely based on visual information, without explicitly modeling reasoning of social content beyond the observed joint actions, the recovery of a socially-relevant dimension is intriguing. Note that this dimension only emerges after introducing the set V features with temporal shift between movement of the two actors. This dimension may be strongly associated with the degree of synchronization of body movements between two actors. For example, the friendly impression of the joint action of "walking towards each other to give a friendly high-five greeting" results from the harmonious body movements created by two actors in cooperation. In contrast, the perceived aggression of the joint action of "a person holding an object moves toward the other actor who follows with an avoiding move" results from the temporal relation of movements between the two actors.

## Conclusions

The experiment reported in this paper showed that human interactivity ratings provide a tool to gauge the psychological space for representing joint actions. A similar approach has been employed in previous research on perceiving emotion from arm movements in individual actions (Pollick et al., 2001). Our study shows that humans represent major two dimensions of joint actions, one concerning visual motion information and the other concerning social aspects of joint actions. To further understand the representational space of joint actions, we assessed the role of critical features of coordinated movements in signaling human interactions. Research on recognition of individual actions has identified critical features involved in the processing of action stimuli (Casile & Giese, 2005; van Boxtel & Lu, 2015). However, few previous studies have systematically examined critical features of coordinated movements in joint actions. Accordingly, the present study fills an important gap in research on action perception.

Humans can perceive activities jointly performed by two actors from very impoverished stimuli such as point-light displays (actions depicted by discrete joints in a motion sequence). Observers can identify whether actors interact in a meaningful way to achieve a shared outcome involving a change in the environment or the fulfillment of a social goal. An important question concerns how the visual system can generalize action perception to point-light stimuli, which are rarely observed in the visual world. The present findings suggest that a robust ability to identify joint actions may involve the extraction of critical features of coordinated body movements between two actors, and making relational connections between these motion features based on inter-actor correlations. Our findings shed light on how humans achieve efficient detection of meaningful inter-personal interactions in complex scenes, paving the way for a deeper understanding of social cognition.

## Acknowledgments

This research was funded by a NSF grant BCS-1353391 to HL and DARPA MSEE project FA 8650-11-1-7149 and ONR MURI project N00014-16-1-2007 for SZ.

## References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-349.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). he autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders,*, *31*, 5-17.

Carpenter, M. (2009). Just how joint is joint action in infancy? *Topics in Cognitive Science*, *1*(2), 380-392.

Casile, A., & Giese, M. A. (2005). Critical features for the recognition of biological motion. *Journal of Vision*, *5*(4), 6.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *American Journal of Psychology*, *57*(2), 243-259.

Marsh, K. L., Richardson, M. J., & Schmidt, R. C. (2009). Social connection through joint action and interpersonal coordination. *Topics in Cognitive Science*, *1*(2), 320-339.

Pollick, F. E., Paterson, H. M., Bruderlin, A., & Sanford, A. J. (2001). Perceiving affect from arm movement. *Cognition*, *82*(2), B51-61.

Richardson, M. J., Marsh, K. L., & Isenhower, R. W. (2007). Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human Movement Science*, *26*(6), 867-891.

Sebanz, N., Bekkering, H., & Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, *10*(2), 70-76.

Thurman, S., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLOS ONE*, *9*(11), 1-12.

Troje, N. F., & Westhoff, C. (2006). The inversion effect in biological motion perception: evidence for a life detector? *Current Biology*, *16*(8), 821-824.

van Boxtel, J., & Lu, H. (2012). Signature movements lead to efficient search for threatening actions. *PLOS ONE*, *7*(5), e37085.

van Boxtel, J., & Lu, H. (2013). A biological motion toolbox for reading, displaying and manipulating motion capture data in research settings. *Journal of Vision*, *13*(12), 1-16.

van Boxtel, J., & Lu, H. (2015). Joints and their relations as critical features in action discrimination: Evidence from a classification image method. *Journal of Vision*, *15*(1), 1-17.