TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 10 (2018) 225–241

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12313

This article is part of the topic "Best of Papers from the Cognitive Science Society Annual Conference," Wayne D. Gray (Topic Editor). For a full listing of topic papers, see http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1756-8765/earlyview

Perception of Human Interaction Based on Motion Trajectories: From Aerial Videos to Decontextualized Animations

Tianmin Shu,[a†] Yujia Peng,[b†] Lifeng Fan,[a] Hongjing Lu,[a,b] Song-Chun Zhu[a,c]

[a]*Department of Statistics, University of California, Los Angeles*
[b]*Department of Psychology, University of California, Los Angeles*
[c]*Department of Computer Science, University of California, Los Angeles*

Received 18 September 2017; accepted 6 October 2017

Abstract

People are adept at perceiving interactions from movements of simple shapes, but the underlying mechanism remains unknown. Previous studies have often used object movements defined by experimenters. The present study used aerial videos recorded by drones in a real-life environment to generate decontextualized motion stimuli. Motion trajectories of displayed elements were the only visual input. We measured human judgments of interactiveness between two moving elements and the dynamic change in such judgments over time. A hierarchical model was developed to account for human performance in this task. The model represents interactivity using latent variables and learns the distribution of critical movement features that signal potential interactivity. The model provides a good fit to human judgments and can also be generalized to the original Heider–Simmel animations (1944). The model can also synthesize decontextualized animations with a controlled degree of interactiveness, providing a viable tool for studying animacy and social perception.

Keywords: Social interaction; Motion; Decontextualized animation; Hierarchical model; Action understanding

Correspondence should be sent to Tianmin Shu, Department of Statistics, University of California, Los Angeles, CA 90095. E-mail: tianmin.shu@ucla.edu.

[†]The first two authors contributed equally.

1. Introduction

People are adept at perceiving goal-directed action and inferring social interaction from movements of simple objects. In their pioneering work, Heider and Simmel (1944) presented video clips showing three simple geometrical shapes moving around and asked human observers to describe what they saw. Almost all observers described the object movements in an anthropomorphic way, reporting a reliable impression of animacy and meaningful social interactions among the geometric shapes displayed in the decontextualized animation. Their results were replicated in other studies using similar videos for both human adults (Oatley & Yuill, 1985; Rimé, Boulanger, Laubin, Richir, & Stroobants, 1985) and preschoolers as young as 5 years old (Springer, Meier, & Berry, 1996).

To study what visual information drives the perception of interaction, Berry, Misovich, Kean, and Baron (1992) generated new Heider–Simmel animations with either the structural aspect or dynamic aspect disrupted. They found that the motion patterns mainly determined the anthropomorphic description of videos. Later studies (Dittrich & Lea, 1994; Gao, Newman, & Scholl, 2009; Gao, McCarthy, & Scholl, 2010; Scholl & Tremoulet, 2000; Tremoulet & Feldman, 2000, 2006) used more controlled stimuli and systematically examined what factors can impact the perception of goal-directed actions in a decontextualized animation. These findings provided converging evidence that the perception of human-like interactions relies on some critical low-level motion cues, such as speed and motion direction. However, it remains unclear how the human visual system combines motion cues from different objects to infer interpersonal interactivity in the absence of any context cues.

To address this fundamental question, Baker, Saxe, and Tenenbaum (2009) developed a Bayesian model to reason about the intentions of an agent when moving in maze-like environments of the sort used by Heider and Simmel (1944). Other studies (Baker, 2012; Baker, Goodman, & Tenenbaum, 2008; Baker, Saxe, & Tenenbaum, 2011; Sadilek & Kautz, 2012; Ullman, Baker, Macindoe, Evans, Goodman, & Tenenbaum, 2009) developed similar models that could be generalized to situations with multiple agents and different contexts. These modeling studies illustrate the potential fruitfulness of using a Bayesian approach as a principled framework for modeling human interaction shown in decontextualized animations. However, these models have been limited to experimenter-defined movements and by computational constraints imposed by the modelers for particular application domains.

In daily life, humans rarely observe Heider–Simmel-type animations. Although examining inferences about human interactions in videos of daily-life activities would be ecologically natural, challenges arise. Human interactions are usually accompanied by rich context information, such as language, body gestures, moving trajectories of multiple agents, and backgrounds in the environment. Hence, the complexity of information may make it difficult to pin down what critical characteristics in the input determine human judgments.

To address this problem, we used aerial video and employed advanced computer vision algorithms to generate experimental stimuli that were rigorously controlled but rooted in real-life situations. As an example, imagine that you are watching a surveillance video recorded by a drone from a bird's eye view, as shown in Fig. 1. In such aerial videos, changes in human body postures can barely be seen, and the primary visual cues are the noisy movement trajectories of each person in the scene. This situation is analogous to the experimental stimuli used in Heider and Simmel animations, but the trajectories of each entity are directly based on real-life human movements. Another advantage of using aerial videos is that they provide a window to examine whether a model trained with real-life motions can generalize its learned knowledge to interpret decontextualized movements of geometric shapes, without prior exposures. Such generalizability emulates humans' irresistible and automatic impressions when viewing the Heider–Simmel animations for the first time. If the generalization is successful, the cues used by the model in learning can shed light on the mechanisms underlying the human ability to recover the causal and social structure of the world from the visual inputs.

In the present study, we aimed to use real-life aerial videos to generate Heider–Simmel-type decontextualized animations and to assess how human judgments of interactivity emerge over time. We employed decontextualized animations generated from the aerial videos to measure how well humans make online judgments about interpersonal interactions and to gauge what visual cues determine the dynamic changes in human judgments. To account for human performance, we developed a hierarchical model with hidden layers. The model aimed to learn the representations of critical movement patterns that signal potential interactivity between agents. Furthermore, we assessed whether the learning component in the model can be generalized to the original animations used by Heider and Simmel (1944).

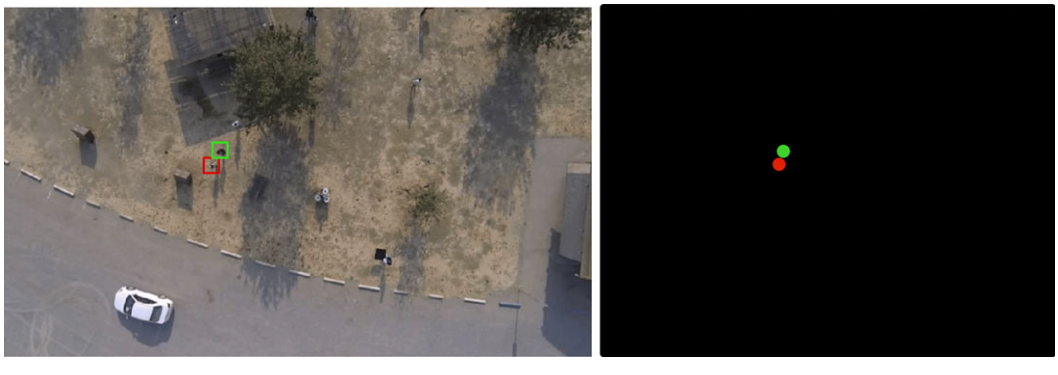

Fig. 1. Stimulus illustration. (Left) An example frame of an aerial video recorded by a drone. Two people were being tracked (framed by red and green boxes). (Right) A sample frame of an experimental trial. The two people being tracked in the aerial video are presented as two dots, one in red and one in green, against a black background. A video demonstration can be viewed on the project website: http://www.stat.ucla.edu/~tianmin.shu/HeiderSimmel/CogSci17.

2. Computational model

We designed a hierarchical model with three layers. As shown in Fig. 2, the first layer (the X layer) estimates spatiotemporal motion patterns within a short period of time. The second layer (the S layer) captures the involvement of various motion fields at different stages of interactivity over a long period by temporally decomposing interactivity into multiple latent subinteractions. The last layer (the Y layer) indicates the presence or absence of interactivity between two agents.

The inputs to the model are motion trajectories of two agents, denoted as $\Gamma_a = \{\mathbf{x}_a^t\}_{t=0,\ldots,T}$, $a = 1, 2$. The position of agent a ($a = 1, 2$) at time t is $\mathbf{x}_a^t = (x, y)$. The total length of the trajectory is T. Using the input of motion trajectories, we can readily compute the velocity sequence of agent a ($a = 1, 2$), i.e., $V_a = \{\mathbf{v}_a^t\}_{t=1,\ldots,T}$, where $\mathbf{v}_a^t = \mathbf{x}_a^t - \mathbf{x}_a^{t-1}$.

To capture the interactivity between two agents based on the observed trajectories of movements, the model builds on two basic components. (a) Interactivity between two agents can be represented by a sequence of latent motion fields, each capturing the relative motion between the two agents who perform meaningful social interactions. (b) Latent motion fields can vary over time, capturing the behavioral change of the agents over a long period of time. The details for quantifying the two key components are presented in the next two subsections.

2.1. Conditional interactive fields

As illustrated in Fig. 3, we use conditional interactive fields (CIFs) to represent how an agent moves with respect to a reference agent. This is analogous to the force fields in physics, where the objects interact with each other through invisible fields (e.g., gravity). To derive the CIFs, we randomly select an agent to be the reference agent, and then

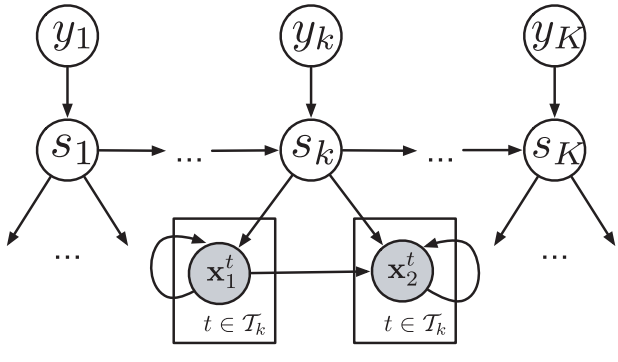

Fig. 2. Illustration of the hierarchical generative model. The solid nodes are observations of motion trajectories of two agents, and the remaining nodes are latent variables constituting the symbolic representation of an interaction; that is, the original trajectories are coded as a sequence of subinteractions S and interaction labels Y.

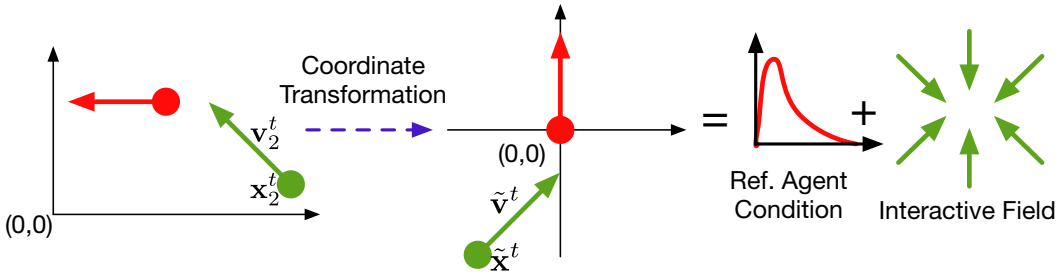

Fig. 3. Illustration of a conditional interactive field: after a coordinate transformation with respect to the reference agent, we model the expected relative motion pattern $\tilde{\mathbf{x}}^t$ and $\tilde{\mathbf{v}}^t$ conditioned on the reference agent's motion.

model the partner agent's movement by estimating a vector field of the relative motion conditioned on a specific distribution of the reference agent's motion.

To ensure that the fields are orientation invariant, we perform a coordinate transformation, as Fig. 3 illustrates. At each time point t, the transformed position of the reference agent is always located at (0, 0), and its transformed velocity direction is always pointed to the norm of the upward vertical direction. Consequently, the position and velocity of the second agent after the transformation, i.e., $\tilde{\Gamma} = \{\tilde{\mathbf{x}}^t\}_{t=0,\ldots,T}$ and $\tilde{V} = \{\tilde{\mathbf{v}}^t\}_{t=1,\ldots,T}$, can be used to model the relative motion.

A subinteraction s corresponds to interactivity in a relatively short time sharing consistent motion patterns, for example, approaching, walking together, standing together. The model can infer its CIF using a potential function $U(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t)$, where the first two variables $(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t)$ are used to model the relative motion as defined in the last paragraph and $\mathbf{v}_1^t$ is the reference agent's motion. The potential function is defined to yield the lowest potential value if the motion pattern fits the characteristics of s the best. In this way, the model considers the agents more likely to be interactive if the agents are moving in a specific way that can minimize the potential energy with respect to certain potential fields.

2.2. Temporal parsing by latent subinteractions

We assume that a long interactive sequence can be decomposed into several distinct subinteractions each with a different CIF. For example, when observing that two people walk toward each other, shake hands, and walk together, this long sequence can be segmented into three distinct subinteractions. We represent meaningful interactivity as a sequence of latent subinteractions $S = \{s_k\}_{k=1,\ldots,K}$, where a latent subinteraction determines the category of the CIF involved in a time interval $\mathcal{T}_k = \{t : t_k^1 \leq t \leq t_k^2\}$, such that $s^t = s_k, \forall t \in \mathcal{T}_k$. s_k is the subinteraction label in the k-th interval representing the consistent interactivity of two agents in the relatively short interval. Fig. 4 illustrates the temporal parsing.

In each interval k, we define an interaction label $y_k \in \{0, 1\}$ to indicate the absence or presence of interactivity between the two agents. The interaction labels also constitute a

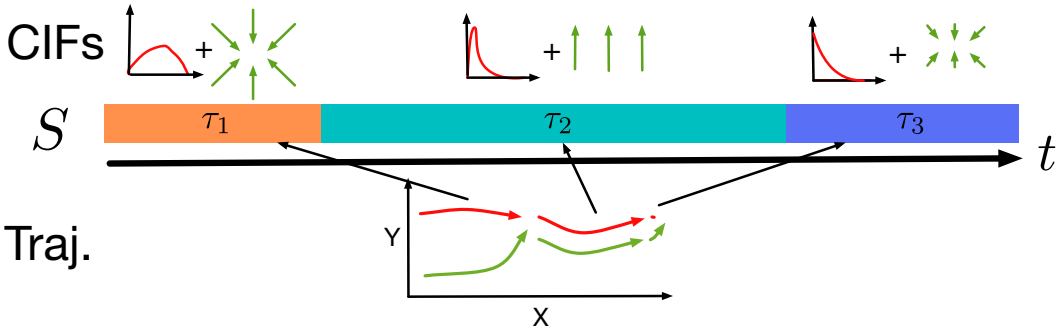

Fig. 4. Temporal parsing by S (middle). The top demonstrates the change of conditional interactive fields (CIF) in subinteractions as the interaction proceeds. The bottom indicates the change of interactive behaviors in terms of motion trajectories. The colored bars in the middle depict the types of the subinteractions.

sequence $Y = \{y^t\}_{t=1,\ldots,T}$. We have $y^t = y_k$, $\forall t \in \mathcal{T}_k$, where y_k denotes the interaction label in an interval $\mathcal{T}_k$.

3. Model formulation

Given the input of motion trajectories Γ as defined in the above section, the model infers the posterior distribution of the latent variables S and Y using a Bayesian framework,

$$p(S, Y|\Gamma) \propto \underbrace{P(\Gamma|S, Y)}_{\text{likelihood}} \times \underbrace{P(S|Y)}_{\text{sub int. prior}} \times \underbrace{P(Y)}_{\text{int. prior}}. \tag{1}$$

The likelihood assesses how well the motion fields represented as a set of subinteractions CIFs can account for relative motion observed in the video input, the spatial density of the relative position, and the observed motion of the reference agent:

$$p(\Gamma|S, Y) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t | s^t = s_k, y^t = y_k), \tag{2}$$

where the individual likelihood terms are defined by potential functions:

$$\log p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t | s^t = s_k, y^t = y_k) \propto -U(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t | s_k, y_k). \tag{3}$$

Here, we assume that the potential function depends on the latent variables s_k and y_k to account for the variability in the motion patterns of different subinteractions and to differentiate interactive motion from non-interactive motion. Equation 3 also ensures that the expected interactive motion trajectories will move in the direction that minimizes the

potential energy. The Appendix A provides details of the exact definition of the potential function.

We model the prior term of subinteractions $P(S|Y)$ using two independent components, (a) the duration of each subinteraction and (b) the transition probability between two consecutive subinteractions, as follows:

$$p(S|Y) = \prod_{k=1}^{K} \underbrace{p(|\mathcal{T}_k||s_k, y_k)}_{\text{duration}} \prod_{k=2}^{K} \underbrace{p(s_k|s_{k-1}, y_k)}_{\text{transition}}. \tag{4}$$

When $y_k = 1$, the two terms follow a log-normal distribution and a categorical distribution, respectively; when $y_k = 0$, uniform distributions are used for the two terms instead.

Finally, we use a Bernoulli distribution to model the prior term of interactions $P(Y)$,

$$p(Y) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(y^t = y_k) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} \rho^{y^t} (1 - \rho)^{1 - y^t}. \tag{5}$$

The details of model implementation regarding inference and learning is included in Appendix B and C sections.

4. Model simulation results

We trained the model using two sets of training data, the UCLA aerial event dataset (Shu, Xie, Rothrock, Todorovic, & Zhu, 2015), and the Heider–Simmel animation dataset.

4.1. Training with aerial videos

In the UCLA aerial event dataset collected by Shu et al. (2015), about 20 people performed some group activities in two scenes (a park or a parking lot), such as group touring, queuing in front of a vending machine, or playing Frisbee. People's trajectories and their activities are manually annotated. The dataset is available at http://www.stat.ucla.edu/tianmin.shu/AerialVideo/AerialVideo.html.

One advantage of using aerial videos to generate decontextualized animations is that the technique provides sufficient training stimuli to enable the learning of representations of critical movement patterns that signal potential interactivity between agents. We selected training videos including interactivity from the database, so that the two agents always interact with each other in all training stimuli. Thus, for any training video, $y^t = 1$, $\forall t = 1, \ldots, T$. During the training phase, we excluded the examples used in human experiments. In total, there were 131 training instances.

In the implementation, we manually define the maximum number of subinteraction categories to be 15 in our full model (i.e., $|\mathcal{S}| = 15$), which is over-complete for our

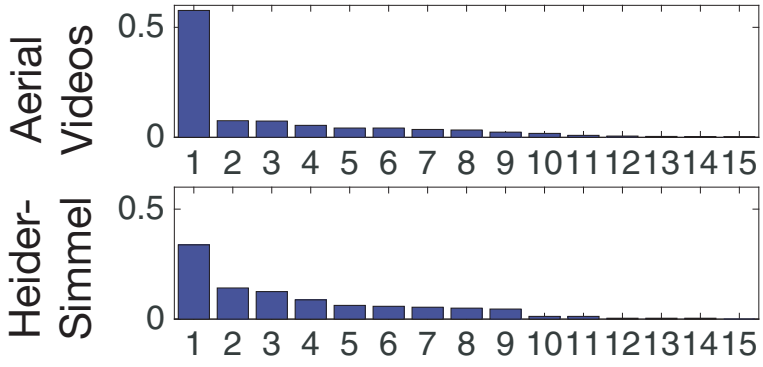

Fig. 5. The frequencies of learned conditional interactive fields (CIFs) with the training data generated from aerial videos (top) and the Heider–Simmel movie (bottom). The numbers on the x axis indicate the IDs of CIFs, ranked according to the occurrence frequency in the training data.

training data according to learning (low frequency in the tail of Fig. 5). With simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983), Gibbs sampling converges within 20 sweeps (where a sweep is defined as all the latent subinteraction labels being updated once). The frequencies of the top 15 CIFs are highly unbalanced. In fact, the top 10 CIFs account for 83.8% of the subinteractions in the training data. The first row of Fig. 6 provides a visualization of the top 5 CIFs. Each of the top CIFs indicates some different behavioral patterns in the aerial videos. For example, the No.1 CIF signals the approaching behavior that one agent moves toward a reference agent. Interestingly, the converging point of the approaching is not at the center of the location of the reference agent. Instead, the agent heads toward the future location of the reference agent (above-the-

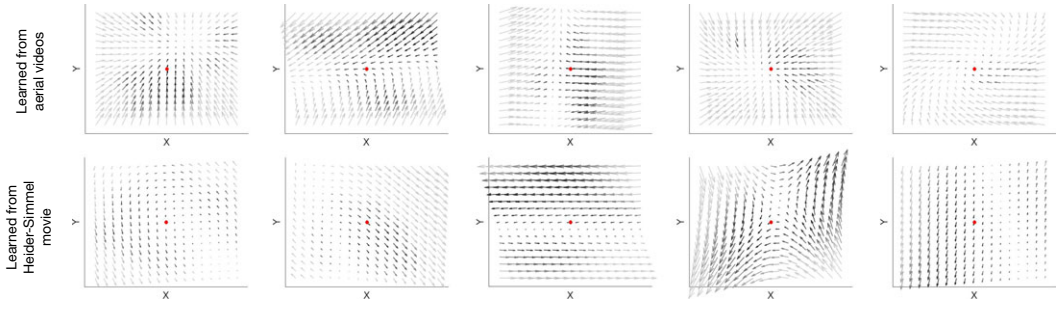

Fig. 6. Interactive fields of the top five frequent conditional interactive fields (CIFs) learned from aerial videos (top) and Heider–Simmel movie (bottom), respectively. In each field, the reference agent (red dot) is at the center of a field; that is, (0, 0), moving toward north; the arrows represent the mean relative motion at different locations and the intensities of the arrows indicate the relative spatial density, which increases from light to dark. We observed a few critical CIFs that signal common interactions from the two simulation results. For instance, in aerial videos, we observed (i) approaching, for example, CIF 1, and (ii) walking in parallel, or following, for example, the lower part of CIF 2. The Heider–Simmel animation revealed additional patterns such as (i) orbiting, for example, CIF 1, and (ii) leaving, for example, CIF 4, (iii) walking-by, for example, CIF 5.

center position in the flow figure), implying that the fundamental characteristic of human interactions is being predictive.

4.2. Training with Heider–Simmel videos

The second dataset was created from the original Heider–Simmel animation (i.e., two triangles and one circle). We extracted the trajectories of the three shapes and thus obtained three pairs of two-agent interactions. We truncated the movie into short clips (about 10 s) to generate a total of 27 videos. The same algorithm was used to train the model with 15 types of CIFs.

The most frequent five CIFs are visualized in the second row of Fig. 6. Clearly, the richer behavior in the Heider–Simmel animation yielded a variety of CIFs with distinct patterns compared to the CIFs learned from aerial videos. For example, the top CIF indicates that one agent moves around the reference agent, a common movement pattern observed in Heider–Simmel animations. The second CIF signals a "run away" movement to avoid the reference agent. The frequencies of CIFs are also more distributed in this dataset, as shown in Fig. 5.

4.3. Generalization: Training with aerial videos and testing with Heider–Simmel videos

We tested how well the model trained with the aerial videos ($|\mathcal{S}| = 15$) can be generalized to a different dataset, the Heider–Simmel animations. This generalization test aims to examine if the critical movement patterns learned from real-life situations can account for perceived interactiveness in laboratory stimuli. Fig. 7 shows the model simulation results for a few Heider–Simmel videos. We notice that the interactiveness ratings predicted by the model vary over time. Such variability is consistent with subjective impressions that the Heider–Simmel animations elicit different degrees of animacy and interactivity at different time points. In addition, most clips in Heider–Simmel animations are rated by the model as having a high probability of being interactive (i.e., mostly

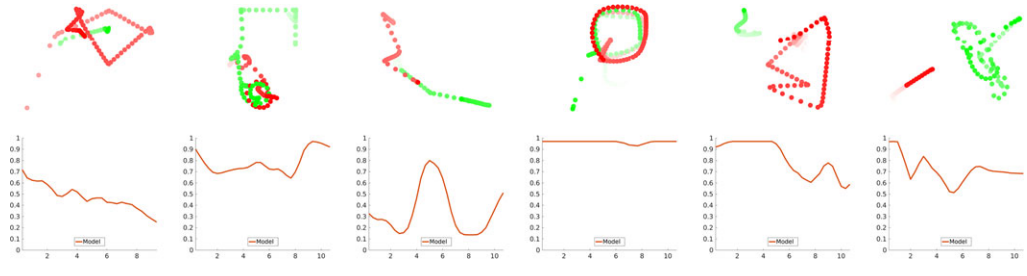

Fig. 7. (Top) Examples of moving trajectories of selected objects in the Heider–Simmel animation dataset. One object is plotted in red and the other one is plotted in green. The intensity of colors increases with time lapse, with darker color representing more recent coordinates. (Bottom) Corresponding online predictions on the example Heider–Simmel videos by our full model ($|\mathcal{S}| = 15$) trained on aerial videos over time (in seconds).

above 0.5), consistent with human observers' impression about the highly animate and interactive behaviors conveyed in the animations. Also, the model was able to give continuous online predictions based on the relative speeds and spatial locations of the two objects. For example, when the two objects approach each other or follow each other, the model yields higher interactive ratings.

The qualitative analysis of the model performance suggests that the model trained with aerial videos shows a certain degree of generalization to the Heider–Simmel animations. However, unsurprisingly, objects in aerial videos share different characteristics of motion patterns from the motions involved in Heider–Simmel animations (as illustrated in the training results of CIFs in Fig. 6). For example, orbiting behavior barely occurs in the aerial video dataset, and accordingly the model yields relatively low interactiveness predictions when observing such behavior, which is relatively common in the Heider–Simmel animations. In the next section, we will report human experiments that can quantitatively assess how well the model can account for human performance.

5. Experiment 1

5.1. Stimuli

Twenty-four interactive stimuli were generated from different pairs of human interactions in aerial videos. We selected two people interacting with each other in each aerial video. We then generated the decontextualized animations by depicting the two people as dots with different colors. The dots' coordinates were first extracted from the aerial videos by human annotators. Note that the two dots were first recentered to localize the midpoint at the center of the screen in the first frame. The coordinates were temporally smoothed by averaging across the adjacent five frames.

Twenty-four non-interactive stimuli were generated by interchanging motion trajectories of two people selected from two irrelevant interactive videos (e.g., the motion of one dot in video 1 recombined with the motion of a dot in video 2). The starting distances between two dots in non-interactive stimuli were kept the same as in the corresponding interactive stimuli.

The duration of stimuli varied from 239 frames to 500 frames (mean frame = 404), corresponding to 15.9 to 33.3 s, with a recording refresh rate of 15 frames per second. The diameters of dots were 1° of visual angle. One dot was displayed in red (1.8 cd/m^2) and the other in green (30 cd/m^2) on a black background (0 cd/m^2). Among the 48 pairs of stimuli, four pairs of actions (two interactive and two non-interactive) were used as practice.

5.2. Participants

Thirty-three participants (M_{age} = 20.4; 18 female) were enrolled from the subject pool at the Department of Psychology, University of California, Los Angeles (UCLA). They

were compensated with course credit. All participants had normal or corrected-to-normal vision.

5.3. Procedures

Participants were seated 35 cm in front of a screen, which had a resolution of 1,024 × 768 and a 60 Hz refresh rate. First, participants were given a cover story: "Imagine that you are working for a company to infer whether two people carry out a social interaction based on their body locations measured by GPS signals. Based on the GPS signal, we generated two dots to indicate the location of the two people being tracked." The task was to determine when the two dots were interacting with each other and when they were not. Participants were asked to make continuous responses across the entire duration of the stimuli. They were to press and hold the left-arrow or right-arrow button for interactive or non-interactive moments, respectively, and to press and hold the down-arrow button if they were unsure. If no button was pressed for more than one second, participants received a 500 Hz beep as a warning.

Participants were presented with four trials of practice at the beginning of the session to familiarize them with the task. Next, 44 trials of test stimuli were presented. The order of trials was randomized for each participant. No feedback was presented on any of the trials. The experiment lasted for about 30 min in total.

5.4. Results

Interactive, unsure, and non-interactive responses were coded as 1, 0.5, and 0, respectively. Frames with no responses were removed from the comparison. Human responses are shown in Fig. 8. A paired-sample t-test revealed that the average ratings of

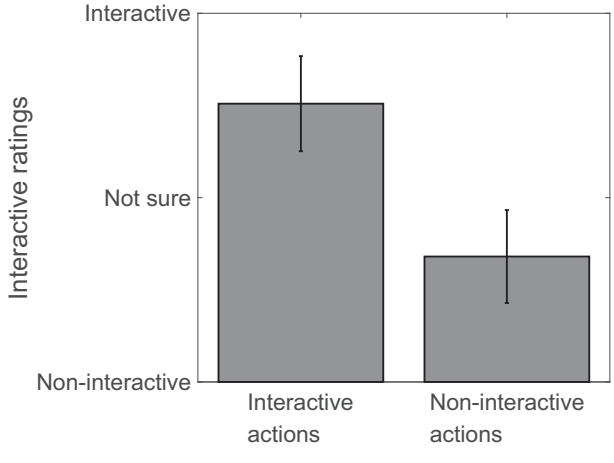

Fig. 8. Mean ratings of the interactive versus non-interactive actions in the experiment 1. Error bars indicate ±1 *SEM*.

non-interactive actions ($M = 0.34$, $SD = 0.13$) were significantly lower than interactive actions ($M = 0.75$, $SD = 0.13$), $t(32) = 13.29$, $p < .001$. This finding indicates that human observers are able to discriminate interactivity based on decontextualized animations generated from the real-life aerial videos.

To compare the model predictions with human continuous judgments, we computed the average human ratings and ran the model to simulate online predictions of subinteraction and interaction labels on the testing videos (excluding the ones in the validation set). Specifically, we used Eq. B3 to compute the probability of two agents being interactive with each other at any time point t. The model simulation used the hyper-parameters $\rho = 10^{-11}$ and $\sigma_0 = 1.26$.

Table 1 summarizes the Pearson correlation coefficient r and root-mean-square error (*RMSE*) between the model predictions and the human ratings using aerial videos as training data. We compared our hierarchical model with two baseline models: (a) hidden Markov model, where the latent variables s^t and y^t only depend on their preceding variables s^{t-1} and y^{t-1}; (b) a model with only one type of subinteraction. Both models yielded poorer fits to human judgments (i.e., lower correlation and higher *RMSE*) than the hierarchical model. In addition, we changed the number of sub-interaction categories to examine how sensitive our model is to this parameter. The results clearly show that (a) only using one type of subinteraction provides reasonably good results, $r = .855$, and (b) by increasing the number of subinteractions $|\mathcal{S}|$, the fits to human ratings were further improved until reaching a plateau with a sufficiently large number of subinteractions.

Fig. 9 shows results for a few videos, with both model predictions and human ratings. The model predictions accounted for human ratings quite well in most cases. However, the model predictions were slightly higher than the average human ratings, which may be due to the lack of negative examples in the training phase. We also observed high standard deviations in human responses, indicating large variability of the online prediction task for every single frame in a dynamic animation. In general, the difference between our model's predictions and human responses is seldom larger than one standard deviation relative to human responses.

We also used the model trained from the Heider–Simmel animation and tested it on the stimuli generated from the aerial videos. This procedure yielded a correlation of 0.640 and *RMSE* of 0.227. The reduced fit for this simulation indicates the discrepancy between moving patterns of the two types of training datasets. The CIFs learned from one dataset may be limited in generalization to the other dataset.

Table 1
The quantitative results of all methods in Experiment 1 using aerial videos as training data

			Hierarchical Model		
Method	Hidden Markov Model	One-Interaction	$\lvert\mathcal{S}\rvert = 5$	$\lvert\mathcal{S}\rvert = 10$	$\lvert\mathcal{S}\rvert = 15$
r	.739	.855	.882	.911	.921
RMSE	0.277	0.165	0.158	0.139	0.134

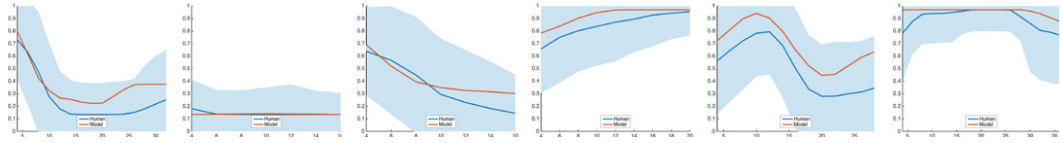

Fig. 9. Comparison of online predictions by our full model trained on aerial videos ($|\mathcal{S}| = 15$) (orange) and humans (blue) over time (in seconds) on testing aerial videos. The shaded areas show the standard deviations of human responses at each moment.

6. Experiment 2

One advantage of developing a generative model is that it enables the synthesis of new videos by Eqs. B4 and B5, based on randomly sampled initial positions of the two agents ($\mathbf{x}_1^0$, $\mathbf{x}_2^0$) and the first subinteraction s^1. By setting the interaction labels to be 1 or 0, the synthesized stimuli can be controlled to vary the degree of interactiveness. In Experiment 2, we aimed to use the model to synthesize new animations and see if interactiveness can be accurately perceived by human observers.

We used the model trained on aerial videos to synthesize 10 interactive and 10 non-interactive animation clips. Seventeen participants were enrolled from the subject pool at UCLA. The procedure of Experiment 2 was similar to that of Experiment 1. The 20 synthesized videos were presented to human observers in random orders. The task was to press one of the two buttons at the end of the action to judge if the two dots were interacting or not.

The interactiveness between the two agents in the synthesized videos was judged accurately by human observers, with the average ratings of the synthesized non-interactive actions ($M = 0.15$, $SD = 0.15$) significantly lower than the synthesized interactive actions ($M = 0.85$, $SD = 0.20$), $t(16) = 14.00$, $p < .001$. The model prediction of a whole video is set to be the average predictions of Eq. B3. The correlation between model predictions and average human responses was high, 0.94. The results suggested that humans reliably perceived interactiveness from the synthesized stimuli and were sensitive to model-controlled degree of interactivity.

7. Discussion

In this paper, we examined human perception of social interactions using decontextualized animations based on movement trajectories recorded in aerial videos of a real-life environment, as well as Heider–Simmel-type animations. The proposed hierarchical model built on two key components: CIFs of subinteractions, and temporal parsing of interactivity. The model fits human judgments of interactiveness well, and it suggests potential mechanisms underlying our understanding of meaningful human interactions. Human interactions can be decomposed into subinteractions such as approaching, walking in parallel, or standing still in close proximity. Based on the transition probabilities and

the duration of subcomponents, humans are able to make inferences about how likely the two people are interacting.

Our study indicates that rapid judgments on human interactivity can be elicited by the detection of critical visual features such as CIFs, without the involvement of a high-level reasoning system. The fairly fast, automatic, irresistible, and highly stimulus-driven impressions about animacy and interactivity are largely perceptual in nature. This result is consistent with the literature on causal perception (Peng, Thurman, & Lu, 2017; Scholl & Tremoulet, 2000) and biological motion perception (Johansson, 1973; Su, van Boxtel, & Lu, 2016; Thurman & Lu, 2013, 2014; van Boxtel & Lu, 2011, 2012). Hence, the detection of interactivity between agents is likely to be processed as in the proposed model without the explicit modeling of intention and goals. This process is efficient, but not sufficient to address questions such as why and how the interactions are carried out between the agents. When these questions are important for a particular task in the social context, the reasoning system and the theory-of-mind system will be called upon after the perception of interactivity has been signaled. Future work should focus on the interplay between the two general systems involved in perception and in inference of human interactions.

The model provides a general framework and can be extended to include hidden intentions and goals. By modifying the potential function in the model, the computational framework can be applied to more sophisticated recognition and understanding of social behavioral patterns. While previous work has focused on actions of individuals based on detecting local spatial-temporal features embedded in videos (Dollár, Rabaud, Cottrell, & Belongie, 2005), the current work can deal with multi-agent interactions. Understanding the relation between agents could facilitate the recognition of individual behaviors by putting single actions into meaningful social contexts. The present model could be further improved to enhance its flexibility and broaden its applications. The parametric linear design of CIFs provides computational efficiency, and temporally parsing an interaction into multiple subinteractions enhances the linearity in each subinteraction. However, this design may not be as flexible as non-parametric or non-linear models, such as a Gaussian process. In addition, the current model is only based on visual motion cues. The model could be enhanced by incorporating a cognitive mechanism (e.g., a theory-of-mind framework) to enable explicit inference of intentions.

Acknowledgments

This research was funded by an NSF grant BCS-1353391 to HL, CSC scholarship to YP, and DARPA MSEE project FA 8650-11-1-7149 and ONR MURI project N00014-16-1-2007 for SZ. This work has been published as a conference proceeding paper in 2017, and it received the computational model prize in perception from the Cognitive Science Society.

References

Baker, C. L. (2012). Bayesian theory of mind: Modeling human reasoning about beliefs, desires, goals, and social relations. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Baker, C. L., Goodman, N. D., & Tenenbaum, J. B. (2008). Theory-based social goal inference. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1447–1452). Austin, TX: Cognitive Science Society.

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

Baker, C. L., Saxe, R., & Tenenbaum, J. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Cognitive Science Society*, *33*, 2469–2474.

Berry, D. S., Misovich, S. J., Kean, K. J., & Baron, R. M. (1992). Effects of disruption of structure and motion on perceptions of social causality. *Personality and Social Psychology Bulletin*, *18*(2), 237–244.

Dittrich, W. H., & Lea, S. E. (1994). Visual perception of intentional motion. *Perception*, *23*(3), 253–268.

Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 10th IEEE International Conference on Computer Vision*. Beijing, China.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154–179.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*, 1845–1853.

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

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220* (4598), 671–680.

Oatley, K., & Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *British Journal of Social Psychology*, *24*(2), 115–124.

Peng, Y., Thurman, S., & Lu, H. (2017). Causal action: A fundamental constraint on perception and inference about body movements. *Psychological Science*, *28*(6), 798–807. https://doi.org/10.1177/0956797617697739

Rimé, B., Boulanger, B., Laubin, P., Richir, M., & Stroobants, K. (1985). The perception of interpersonal emotions originated by patterns of movement. *Motivation and Emotion*, *9*(3), 241–260.

Sadilek, A., & Kautz, H. (2012). Location-based reasoning about complex multi-agent behavior. *Journal of Artificial Intelligence Research*, *43*, 87–133.

Scholl, B. J., & Tremoulet, R. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4* (8), 299–309.

Shu, T., Xie, D., Rothrock, B., Todorovic, S., & Zhu, S.-C. (2015). Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA.

Springer, K., Meier, J. A., & Berry, D. S. (1996). Nonverbal bases of social perception: Developmental change in sensitivity to patterns of motion that reveal interpersonal events. *Journal of Nonverbal Behavior*, *20*(4), 199–211.

Su, J., van Boxtel, J. J., & Lu, H. (2016). Social interactions receive priority to conscious perception. *PloS One*, *11*(8), e0160468.

Thurman, S. M., & Lu, H. (2013). Physical and biological constraints govern perceived animacy of scrambled human forms. *Psychological Science*, *24*(7), 1133–1141.

Thurman, S. M., & Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PloS One*, *9*(11), e112539.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943–951.

Tremoulet, P. D., & Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & Pyschophysics*, *68*(6), 1047–1058.

Ullman, T., Baker, C. L., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. B. (2009). Help or hinder: Bayesian models of social goal inference. In *Proceedings of the 22nd Advances in Neural Information Processing Systems*. Vancouver, BC, Canada.

van Boxtel, J. J., & Lu, H. (2011). Visual search by action category. *Journal of Vision*, *11*(7), 19–19.

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

Appendix A: Formulation

We define the potential function in Eq. 3 as

$$U(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t | s^t = s_k, y^t = y_k) = \mathbf{w}_{s_k,y_k}^\top \phi(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t) + \beta_{s_k,y_k}, \tag{A1}$$

where $\phi(\tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t, \mathbf{v}_1^t) = [\tilde{\mathbf{x}}^{t\top}, \tilde{\mathbf{v}}^{t\top}, \mathbf{v}_1^{t\top}, \tilde{\mathbf{x}}^{t\top}\tilde{\mathbf{v}}^t, \|\tilde{\mathbf{x}}^t\|, \|\tilde{\mathbf{v}}^t\|, \|\mathbf{v}_1^t\|]^\top$ is the motion feature vector used to characterize the potential field, $\mathbf{w}_{s_k,y_k}$ and β_{s_k,y_k} are coefficients of the potential function learned for the specific latent variables s_k and y_k. There are certainly other ways to specify the potential function taking more motion patterns into account, such as acceleration, environment around the agents, and other possible factors of interest.

Appendix B: Inference and prediction

The model infers the current status of latent variables and produces an online prediction of future trajectories. Inference and prediction are performed for each time point from 1 to T sequentially (rather than offline prediction, which gives the labels after watching the entire video).

We denote trajectories from 0 to t as $\Gamma_{0:t}$, and the subinteractions from 1 to $t-1$ as $S_{1:t-1}$. Without loss of generality, we assume there are K subinteractions in $S_{1:t-1}$ with $\mathcal{T}_K$ being the last interval and $s^{t-1} = s_K$. We first infer s^t under the assumption of interaction (i.e., $y^t = 1$) by maximizing

$$p(s^t | \Gamma_{0:t}, S_{1:t-1}, y^t) \propto p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, v_1^t | s^t, y^t) p(s^t | S_{1:t-1}, y^t), \tag{B1}$$

where,

$$p(s^t | S_{1:t-1}, y^t) = \begin{cases} p(\tau \geq |\mathcal{T}_k| + 1 | s^t = s^{t-1}, y^t) & \text{if } s^t = s^{t-1} \\ p(\tau \geq 1 | s^t, y^t) p(s^t | s^{t-1}) & \text{otherwise} \end{cases} \tag{B2}$$

Then the posterior probability of $y^t = 1$ given $s^t \in \mathcal{S}$ is defined as

$$p(y^t|s^t, \Gamma_{0:t}, S_{1:t-1}) \propto p(s^t|\Gamma_{0:t}, S_{1:t-1}, y^t)p(y^t), \tag{B3}$$

This computation makes it possible to perform the following inferences and online prediction: (a) we maximize Eq. B1 to obtain the optimal s^t; (b) we use Eq. B3 to compute the posterior probability of two agents being interactive at t under the CIF of s^t as an approximation of the judgment of interaction/non-interaction provided by human observers; (c) the model can synthesize new trajectories using the following computation,

$$s^{t+1} \sim p(s^{t+1}|S_{1:t}, y^{t+1}), \tag{B4}$$

$$\mathbf{x}_1^{t+1}, \mathbf{x}_2^{t+1} \sim p(\tilde{\mathbf{x}}^{t+1}, \tilde{\mathbf{v}}^{t+1}, v_1^{t+1}|s^{t+1}, y^{t+1}), \tag{B5}$$

where $\tilde{\mathbf{v}}^{t+1}$, $\tilde{\mathbf{x}}^{t+1}$, and v_1^{t+1} are given by $\mathbf{x}_1^t$, $\mathbf{x}_1^{t+1}$, $\mathbf{x}_2^t$ and $\mathbf{x}_2^{t+1}$. By setting $y^{t+1} = 1$ or $y^{t+1} = 0$ in Eqs. B4 and B5, we may synthesize interactive or non-interactive motion trajectories respectively.

Appendix C: Learning

To train the model, we used Gibbs sampling to find the S that maximizes the joint probability $P(Y, S, \Gamma)$. The implementation details are summarized below:

- Step 0: To initialize S, we first construct a feature vector for each time t (see the Appendix A). K-means clustering is then conducted to obtain the initial $\{s^t\}$, which also gives us the subinteraction parsing S after merging the same consecutive s^t.
- Step 1: At each time point t of every training video, we update its subinteraction label s^t by

$$s^t \sim p(\Gamma|S_{-t} \cup \{s^t\}, Y)p(S_{-t} \cup \{s^t\}|Y), \tag{C1}$$

where S_{-t} is the subinteraction temporal parsing excluding time t, and $S_{-t} \cup \{s^t\}$ is a new subinteraction sequence after adding the subinteraction at t. Note that Y is always fixed in the procedure; thus, we do not need $p(Y)$ term for sampling purpose.

- Step 2: If S does not change anymore, go to next step; otherwise, repeat step 1.
- Step 3: Since we do not include the non-interactive videos in the training set, we selected 22 videos in the first human experiment (a mixture of interactive and non-interactive videos) as a validation set to estimate coefficients of the potential functions under $y = 0$ by maximizing the correlation between the model prediction of Eq. B3 and the average human responses in the validation set. To simplify the search, we assume all potential functions under $y = 0$ share the same coefficients across all latent subinteractions.