Inferring Human Interaction from Motion Trajectories in Aerial Videos

Tianmin Shu * **Yujia Peng** * **Lifeng Fan** **Hongjing Lu** **Song-Chun Zhu**
{tianmin.shu, yjpeng, lfan}@ucla.edu hongjing@ucla.edu sczhu@stat.ucla.edu
Department of Psychology and Statistics, University of California, Los Angeles, USA

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 of such judgments over time. A hierarchical model was developed to account for human performance in this task, which 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 controlled degree of interactiveness, providing a viable tool for studying animacy and social perception.

Keywords: social interaction; motion; decontextualized animation; hierarchical model; action understanding

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 interaction among the geometric shapes displayed in the decontextualized animation.

Later studies (Dittrich & Lea, 1994; Scholl & Tremoulet, 2000; Tremoulet & Feldman, 2000, 2006; Gao, Newman, & Scholl, 2009; Gao, McCarthy, & Scholl, 2010) used more controlled stimuli and systematically examined what factors can impact the perception of goal-directed actions in a decontextualized animation. The results 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, Goodman, & Tenenbaum, 2008; Ullman et al., 2009; Baker, 2012) 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.

Figure 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, in a black background. A video demonstration can be viewed on the project website: `http://www.stat.ucla.edu/~tianmin.shu/HeiderSimmel/CogSci17`

The present study aims to generate Heider-Simmel-type decontextualized animations using real-life videos of visual scenes. As a naturalistic 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's studies, but the trajectories of each entity are directly based on real-life human movements.

In the present study, we first used real-life aerial videos to generate decontextualized animations and to assess how human judgments of interactivity emerge over time. We developed a hierarchical model to account for human performance. One advantage of using aerial videos to generate decontextualized animations is that the technique provides sufficient training stimuli to enable the learning of a hierarchical model with hidden layers, which could illuminate 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 by Heider and Simmel (1944).

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.

*These two authors contributed equally.

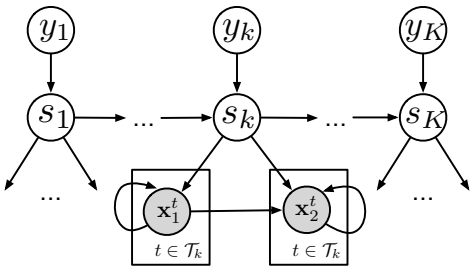

Figure 2: Illustration of the hierarchical generative model. The solid nodes are observations of motion trajectories of two agents, and the remaining nodes are latent variables constituting the symbolic representation of an interaction, i.e., the original trajectories are coded as a sequence of sub-interactions S and interaction labels Y.

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 with latent sub-interactions. The last layer (the Y layer) indicates the presence or absence of interactiveness between two agents.

The inputs to the model are motion trajectories of two agents, denoted as $\Gamma_a = \{\mathbf{x}_a^t\}_{t=0,\cdots,T}$, $a = 1,2$. 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., $V_a = \{\mathbf{v}_a^t\}_{t=1,\cdots,T}$, 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. (1) 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. (2) 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.

Conditional Interactive Fields

As illustrated in Fig. 3, we use conditional interactive fields (CIFs) to model how an agent moves with respect to a reference agent. We randomly select an agent to be the reference agent, and then 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., $\tilde{\Gamma} = \{\tilde{\mathbf{x}}^t\}_{t=0,\cdots,T}$ and $\tilde{V} = \{\tilde{\mathbf{v}}^t\}_{t=1,\cdots,T}$, can be used to model the relative motion.

For a sub-interaction s (interactivity in a relatively short time sharing consistent motion patterns, e.g., approaching, walking together, standing together), we define its CIF as a

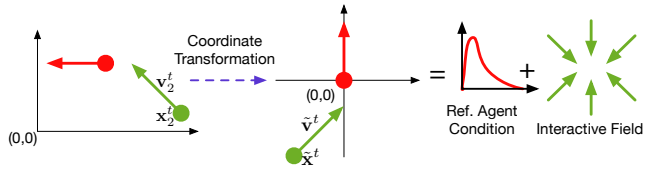

Figure 3: Illustration of a conditional interactive field (CIF): after a coordinate transformation w.r.t. the reference agent, we model the expected relative motion pattern $\tilde{\mathbf{x}}^t$ and $\tilde{\mathbf{v}}^t$ conditioned on the reference agent's motion.

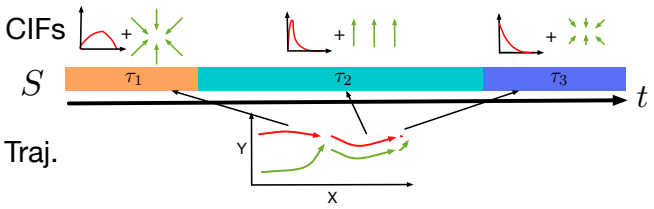

Figure 4: Temporal parsing by S (middle). The top demonstrates the change of CIFs in sub-interactions as the interaction proceeds. The bottom indicates the change of interactive behaviors in terms of motion trajectories. The colored bars in the middle depict the types of the sub-interactions.

linear dynamic system:

$$\tilde{\mathbf{v}}^t \sim \mathcal{N}(A_s\tilde{\mathbf{x}}^t + B_s, \Sigma_s), \tag{1}$$

where A_s, B_s, and $\Sigma_s = diag(\sigma_{s1}^2, \sigma_{s2}^2)$ are the parameters of the Gaussian distribution to be learned for each sub-interaction s. $A_s\tilde{\mathbf{x}}^t + B_s$ can be interpreted as the expected motion at location $\tilde{\mathbf{x}}$ in the field.

Temporal Parsing by Latent Sub-Interactions

We assume that a long interactive sequence can be decomposed into several distinct sub-interactions each with a different CIF. For example, when observing that two people walk towards each other, shake hands and walk together, we can decompose this interactive sequence into three sub-interactions. We represent meaningful interactivity as a sequence of latent sub-interactions $S = \{s_k\}_{k=1,\ldots,K}$, where a latent sub-interaction determines the category of the CIF involved in a time interval $\mathcal{T}_k = \{t : t_k^1 \le t \le t_k^2\}$, such that $s^t = s_k$, $\forall t \in \mathcal{T}_k$. s_k is the sub-interaction label in the k-th interval representing the consistent interactivity of two agents in the 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 sequence $Y = \{y^t\}_{t=1,\cdots,T}$. We have $y^t = y_k$, $\forall t \in \mathcal{T}_k$, where y_k is the interaction label in interval $\mathcal{T}_k$.

Model Formulation

Given the input of motion trajectories Γ, the model infers the posterior distribution of the latent variables S and Y,

$$p(S,Y|\Gamma) \propto \underbrace{P(\Gamma \mid S,Y)}_{\text{likelihood}} \cdot \underbrace{P(S \mid Y)}_{\text{sub int. prior}} \cdot \underbrace{P(Y)}_{\text{int. prior}}. \tag{2}$$

The likelihood assesses how well the motion fields under corresponding CIFs of sub-interactions 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:

$$p(\Gamma \mid S,Y) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t \mid s^t = s_k, y^t = y_k), \quad (3)$$

where

$$\begin{aligned} &p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, \mathbf{v}_1^t \mid s^t = s_k, y^t = y_k) \\ &= \underbrace{p(\tilde{\mathbf{v}}^t \mid \tilde{\mathbf{x}}^t, s_k, y_k)}_{\text{rel. motion}} \cdot \underbrace{p(\tilde{\mathbf{x}}^t \mid s_k, y_k)}_{\text{rel. spatial density}} \cdot \underbrace{p(||\mathbf{v}_1^t|| \mid s_k, y_k)}_{\text{ref. motion}}. \end{aligned} \quad (4)$$

Note that $\mathbf{v}_1^t$ is the reference agent's velocity. When $y_k = 1$, the first term is defined in equation (1), the second term is learned by Gaussian kernel density estimation, and the third term is defined as a Weibull distribution, which is suitable for learning a long-tail distribution of a non-negative variable. When $y_k = 0$, the first term is defined as a Gaussian distribution $\mathcal{N}([0,0]^\top, \Sigma_0 = diag(\sigma_0^2, \sigma_0^2))$, and the remaining two terms are uniform distributions in quantized spaces.

We model the prior term of sub-interactions $P(S|Y)$ using two independent components, i) the duration of each sub-interaction, and ii) the transition probability between two consecutive sub-interactions, as follows:

$$p(S \mid Y) = \prod_{k=1}^{K} \underbrace{p(|\mathcal{T}_k||s_k, y_k)}_{\text{duration}} \prod_{k=2}^{K} \underbrace{p(s_k|s_{k-1}, y_k)}_{\text{transition}}. \quad (5)$$

When $y_k = 1$, the two terms follow a log-normal distribution and a multinomial 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 $P(Y)$,

$$p(Y) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(y^t = y_k) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} \rho^{y^t} (1-\rho)^{1-y^t}. \quad (6)$$

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 sub-interactions from 1 to $t-1$ as $S_{1:t-1}$. Without loss of generality, we assume there are K sub-interactions 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

$$p(s^t \mid \Gamma_{0:t}, S_{1:t-1}, y^t) \propto p(\tilde{\mathbf{v}}^t, \tilde{\mathbf{x}}^t, v_1^t \mid s^t) p(s^t \mid S_{1:t-1}, y^t), \quad (7)$$

where,

$$\begin{aligned} &p(s^t \mid S_{1:t-1}, y^t) \\ &= \begin{cases} p(\tau \geq |\mathcal{T}_k| + 1 \mid s^t = s^{t-1}, y^t) & \text{if } s^t = s^{t-1} \\ p(\tau \geq 1 \mid s^t, y^t) p(s^t | s^{t-1}) & \text{otherwise} \end{cases}. \end{aligned} \quad (8)$$

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

$$p(y^t \mid s^t, \Gamma_{0:t}, S_{1:t-1}) \propto p(s^t \mid \Gamma_{0:t}, S_{1:t-1}, y^t) p(y^t), \quad (9)$$

This computation makes it possible to perform the following inferences and online prediction: i) we maximize (7) to obtain the optimal s^t; ii) we use (9) 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; iii) the model can synthesize new trajectories using the following computation,

$$s^{t+1} \sim p(s^{t+1} \mid S_{1:t}, y^{t+1}), \quad (10)$$

$$\begin{aligned} \mathbf{x}_1^{t+1}, \mathbf{x}_2^{t+1} &\sim p(\mathbf{x}_1^{t+1}, \mathbf{x}_2^{t+1} | \mathbf{x}_1^t, \mathbf{x}_2^t, s^{t+1}, y^{t+1}) \\ &= p(\tilde{\mathbf{v}}^{t+1}, \tilde{\mathbf{x}}^{t+1}, v_1^{t+1} \mid s^{t+1}, y^{t+1}) \end{aligned}, \quad (11)$$

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 $\mathbf{x}_2^{t+1}$, and the last term is defined in (4). By setting $y^{t+1} = 1$ or $y^{t+1} = 0$ in (10) and (11).

Learning

Algorithm

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, i.e., $[||v_1^t||, \tilde{\mathbf{x}}^t, \tilde{\mathbf{v}}^t]^\top$. A K-means clustering is then conducted to obtain the initial $\{s^t\}$, which also gives us the sub-interaction parsing S after merging the same consecutive s^t.
- Step 1: At each time point t of every training video, we update its sub-interaction label s^t by

$$s^t \sim p(\Gamma \mid S_{-t} \cup \{s^t\}, Y) p(S_{-t} \cup \{s^t\} \mid Y), \quad (12)$$

  where S_{-t} is the sub-interaction temporal parsing excluding time t, and $S_{-t} \cup \{s^t\}$ is a new sub-interaction sequence after adding the sub-interaction 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 ρ and σ_0 by maximizing the correlation between the model prediction of (9) and the average human responses in the validation set.

Model Simulation Results

We tested the model using two sets of training data. The first dataset is a UCLA aerial event dataset collected by Shu et al. (2015), in which 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`

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, $y^t = 1, \forall t = 1, \cdots, T$. 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 sub-interaction categories to be 15 in our full model (i.e., $|\mathcal{S}| = 15$), which is over-complete for our training data according to learning (low frequency in the tail of Fig. 6). With simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983), Gibbs sampling converges within 20 sweeps (where a sweep is defined as all the latent sub-interaction labels have been updated once). The frequencies of the top 15 CIFs are highly unbalanced. In fact, the top 10 CIFs account for 83.8% of the sub-interactions in the training data. The first row of Fig. 5 provides a visualization of the top 5 CIFs.

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 3 pairs of two-agent interactions. We truncated the movie into short clips (about 10 seconds) 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. 5. 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. The frequencies of CIFs are also more distributed in this dataset, as shown in Fig. 6.

We observed a few critical CIFs that signal common interactions from the two simulation results. For instance, in aerial videos, we observed i) approaching, e.g., CIF 1 and ii) walking in parallel, or following, e.g., the lower part of CIF 2; the Heider-Simmel animation revealed additional patterns such as i) orbiting, e.g., CIF 1, ii) walking-by, e.g., CIF 5, and iii) leaving, e.g., CIF 4.

Experiment

Stimuli

24 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 re-centered to localize the midpoint at the center of the screen in the first frame. The coordinates were temporally smoothed by averaging across the adjacent 5 frames.

24 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 seconds, 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.

Participants

33 participants (mean age = 20.4; 18 female) were enrolled from the subject pool at the University of California, Los Angeles (UCLA) Department of Psychology. They were compensated with course credit. All participants had normal or corrected-to-normal vision.

Procedures

Participants were seated 35 cm in front of a screen, which had a resolution of 1024×768 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 minutes in total.

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 were shown in Fig. 8 (left). A paired-sample t-test revealed that the average ratings of non-interactive actions (M = 0.34, SD = 0.13) were significantly lower than interactive actions (M = 0.75,

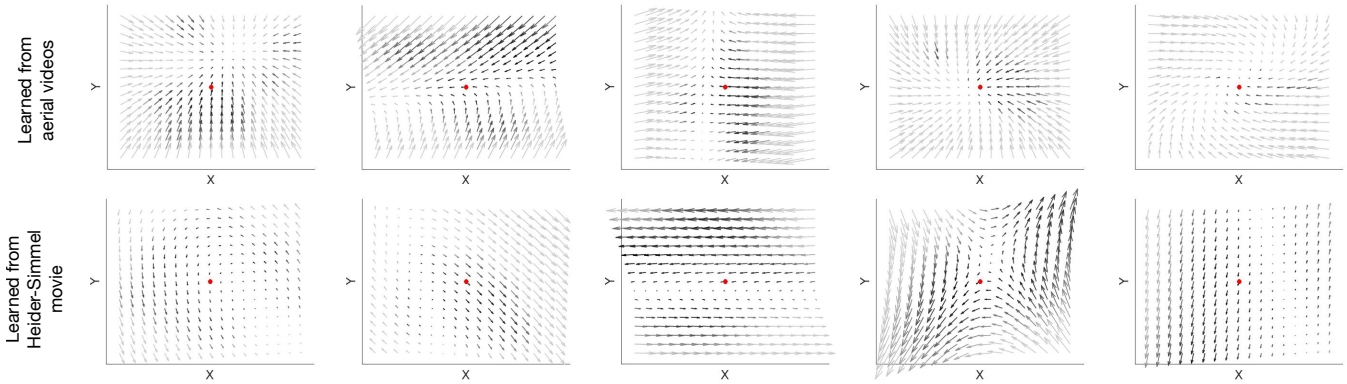

Figure 5: Interactive fields of the top five frequent CIFs learned from aerial videos (top) and Heider-Simmel movie (bottom) respectively. In each field, the reference agent (red dot) is at the center of a field i.e., (0,0), moving towards north; the arrows represent the mean relative motion at different locations and the intensities of the arrows indicate the relative spatial density which increases from light to dark.

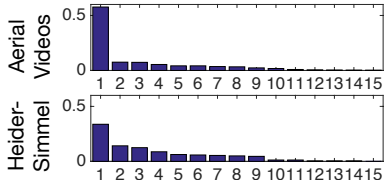

Figure 6: The frequencies of learned CIFs with the training data generated from aerial videos (top) and the Heider-Simmel movie (bottom). The numbers on the x axis indicate the IDs of CIFs, ranked according to the occurrence frequency in the training data.

Method	HMM	One-Interaction	Hierarchical Model		
			$\|\mathcal{S}\| = 5$	$\|\mathcal{S}\| = 10$	$\|\mathcal{S}\| = 15$
r	0.739	0.855	0.882	0.911	0.921
RMSE	0.277	0.165	0.158	0.139	0.134

Table 1: The quantitative results of all methods in experiment 1 using aerial videos as training data.

SD = 0.13), $t(32) = 13.29$, $p < 0.001$. 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 sub-interaction and interaction labels on the testing videos (excluding the ones in the validation set). Specifically, we used (9) 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 compare our hierarchical model with two baseline models: i) Hidden Markov Model (HMM), where the latent variables s^t and y^t only depend on their preceding variables s^{t-1} and y^{t-1}; ii) a model with only one type of sub-interaction. 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 i) only using one type of sub-interaction provides reasonably good results, $r = .855$, and ii) by increasing the number of sub-interactions $|\mathcal{S}|$, the fits to human ratings were further improved until reaching a plateau with a sufficiently large number of sub-interactions.

Fig. 7 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 the 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 are seldom larger than one standard deviation of human responses.

We also tested the model trained from the Heider-Simmel movie on the same testing set (generated from the aerial videos), yielding a correlation of 0.640 and RMSE of 0.227. The reduced fitting result indicates the discrepancy between two types of videos. The CIFs learned from one dataset may be limited in generalization to the other dataset.

One advantage of developing a generative model is that it enables the synthesis of new videos by (10) and (11), based on randomly sampled initial positions of the two agents ($\mathbf{x}_1^0$, $\mathbf{x}_2^0$) and the first sub-interaction s^1. By setting the interaction labels to be 1 or 0, the synthesized stimuli can be controlled to vary the degree of interactiveness. We ran a second experiment using model synthesized animations (10 interactive and 10 non-interactive clips). These synthesized videos were presented to human observers in random orders and the interactive ratings were recorded. The interactiveness between the two agents in the synthesized videos was judged accurately by human observers (mean rating of 0.85 for synthesized interactive clips, and 0.15 for non-interactive clips), suggesting that the model effectively captured the visual features that signal potential interactivity between agents.

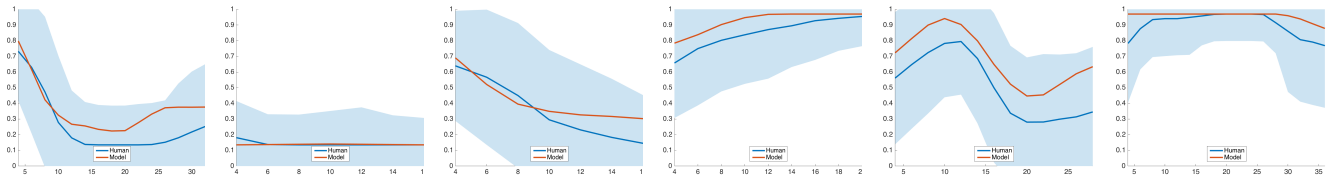

Figure 7: Comparison of online predictions by our full model ($|\mathcal{S}| = 15$) (orange) and humans (blue) over time (in seconds) on testing videos. The shaded areas show the standard deviations of human responses at each moment.

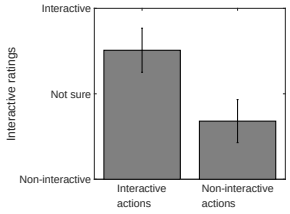

Figure 8: Mean ratings of the interactive versus non-interactive actions in the experiment. Error bars indicate +/- 1 SEM.

Conclusion

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: conditional interactive fields of sub-interactions, and temporal parsing of interactivity. The model fit human judgments of interactiveness well, and suggests potential mechanisms underlying our understanding of meaningful human interactions. Human interactions can be decomposed into sub-interactions such as approaching, walking in parallel, or standing still in close proximity. Based on the transition probabilities and the duration of sub-components, humans are able to make inferences about how likely the two people are interacting.

The model could be extended to be applied to the field of behavioral recognition. 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 interaction. Understanding of the relation between agents could facilitate the recognition of individual behaviors by putting single actions into meaningful social contexts. 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.

Acknowledgement

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. (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 *Proceedings of the thirtieth annual conference of the cognitive science society* (p. 1447-1452).

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

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 *In proceedings of ieee international conference on computer vision workshops* (pp. 65–72).

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

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.

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

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

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 ieee conference on computer vision and pattern recognition.*

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 advances in neural information processing systems* (p. 1874-1882).