ShadowPlay: A Generative Model for Nonverbal Human-Robot Interaction

Eric Meisner
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY
meisne@cs.rpi.edu

Selma Šabanović
Program in Science, Technology and Society
Stanford University
Stanford, CA
selmas@stanford.edu

Volkan Isler
Dept. of Computer Science
University of Minnesota
Minneapolis, MN
isler@cs.umn.edu

Linnda R. Caporael
Dept. of Science and Technology Studies
Rensselaer Polytechnic Institute
Troy, NY
caporl@rpi.edu

Jeff Trinkle
Dept. of Computer Science
Rensselaer Polytechnic Institute
Troy, NY
trink@cs.rpi.edu

ABSTRACT

Humans rely on a finely tuned ability to recognize and adapt to socially relevant patterns in their everyday face-to-face interactions. This allows them to anticipate the actions of others, coordinate their behaviors, and create shared meaning—to communicate. Social robots must likewise be able to recognize and perform relevant social patterns, including interactional synchrony, imitation, and particular sequences of behaviors. We use existing empirical work in the social sciences and observations of human interaction to develop nonverbal interactive capabilities for a robot in the context of shadow puppet play, where people interact through shadows of hands cast against a wall. We show how information theoretic quantities can be used to model interaction between humans and to generate interactive controllers for a robot. Finally, we evaluate the resulting model in an embodied human-robot interaction study. We show the benefit of modeling interaction as a joint process rather than modeling individual agents.

Categories and Subject Descriptors

I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*coherence and coordination*

General Terms

Human Factors Design

HRI'09, March 11–13, 2009, La Jolla, California, USA.

Keywords

Modeling social situations, Interaction synchrony, Nonverbal interaction, Control architecture, Gesture recognition

1. INTRODUCTION

Imagine that you are at a busy pub, trying to order a drink. You strategically wriggle into an available space in front of the bar and focus on the bartenders. As they juggle bottles and orders, your eyes track their movements as if you were watching a tennis match. You lean in, the bar painfully digging into your midriff, vying for physical proximity to the object of your attention. One hand goes up; as they turn in your direction you give a little wave. Your eyes, face, neck, hand, your whole body, follow their movement. And finally—now you know they've noticed you—you are facing each other, making eye contact, smiling, nodding, talking. A connection has been established; you and the bartender are communicating.

This vignette at the bar (and similar ones that occur daily at the restaurant, the library, the post office) illustrates the experience of co-presence—a mutual embodied awareness, a sense of being together with another for shared purposes. In robotics, Breazeal [1] calls this the problem of "embodied discourse," or developing a robot that is able to take part in interaction as an equally proficient participant. To interact, the communicating parties have to establish parity by becoming "coupled" or "linked together by some common base... that puts them on comparable footing, that gives partial mutual access to internal states between them" [2]. Dautenhahn's [3] taxonomy of social robots correspondingly distinguishes between embodied, "situated," and "socially embedded" robots according to their degree of coupling with the social environment. Rather than modeling human and robot as individual agents, we model the joint process that emerges between them.

Recent work in robotics as well as the social sciences suggests that we can construct models of social cognition and behavior by engineering interactive machines and evaluating how humans interact with them [4, 5, 6]. We describe the de-

velopment and validation of a generative model of coupled interaction in the context of dyadic nonverbal interaction—shadow puppet play. Our focus on nonverbal interaction is informed by studies of the foundational nature of interactional synchrony, gestural communication, and imitation to social and communicative development [7, 8], as well as by its fundamental importance to the development of natural human-robot interaction [3, 1, 9, 10, 11].

Our contributions can be summarized as follows. We first describe a method to quantify interaction using human evaluation. Second, we present a perception system which recognizes and codes gestural primitives in real-time. Finally we propose methods for modeling interaction that we use to generate interactive behavior. These models are learned from observing human-human interaction and validated in an embodied human-robot interaction study. Our results show that modeling interaction as a joint process, rather than modeling agents separately, correlates more closely with human evaluations of interactivity.

We provide background discussion on nonverbal interaction and synchrony in section 2. In section 3, we describe how shadow play can be used as a model system for studying nonverbal human-human and human-robot interaction. In section 4, we describe the development of a generative model through observation of people playing shadow puppets aimed at enabling our robot to perceive and interact with gestures. In section 5, we validate the model in an embodied human-robot interaction study. We conclude with a discussion of the relevance of our work for understanding and developing more complex systems in which humans and robots can interact and coordinate their activities.

2. BACKGROUND AND RELATED WORK

Theories of social interaction propose high-level understanding is the result of shared cognition in which the participants learn to coordinate their actions, while low level mechanisms like turn-taking help to coordinate and ground interaction [12, 2]. Humans are able to infer the mental and affective states of others through tone of voice, eye gaze, body language, and other nonverbal behaviors automatically and directly in the course of the interaction. Social engagement triggers an embodied, situated system that is sensitive to recognizing socially relevant patterns in our everyday behavior, such as interaction rhythms, imitation, posture mirroring, and joint attention [13]. Given situational knowledge, this allows us to predict what others will do and coordinate our behaviors. As individuals respond dynamically to bodily movements, postures and facial expressions of others, behavior is used to regulate one's own state and the behavior of other individuals and enables the attunement of intentions among interaction partners. Finally, the combination of rhythmic entrainment, joint attention and coordination makes it possible for a quick blink of an eye to be understood as a socially meaningful "wink."

Nonverbal cues, consisting of bodily movements, articulatory gestures, emotional expressions, and utterances, are instrumental to establishing social presence and mutual awareness. Sustaining participation in interaction involves the reproduction of institutionalized patterns of interaction and habitual practices that are well known to others and can be used to interpret the resulting actions. Interactional synchrony is a pervasive organizing principle of social interaction [14, 15] . Through synchronization, agents establish a common ground for the development of knowledge within the shared interaction, ensuring that what counts for one counts for the other [2]. Studies of infant-caretaker interactions show that coordination is critical to the creation of positive relationships and to the learning of social, cultural, and communicative skills [16]. Condon [7] discusses the importance of the matching of movement timing in smooth, friendly communication, and a higher degree of synchronization is generally regarded as a sign of mutual rapport and involvement [17].

In human-robot interaction studies, studies of nonverbal cues and interaction synchrony have largely focused on sound as the salient stimulus. Kismet [1] relies on rough approximations of turn-taking in conversation. Ogawa's [11] InterRobot humanoid reacts to human speech with nonverbal cues. The humanoid robot Nico [9] synchronizes its drumming to that of another person or a conductor. Another possible reference point for synchrony is bodily motion. Synchronized imitation is applied by Andry [18] as a way for robots to learn new types of motion. Michalowski [10] explores rhythmicity using dance as a social activity. Penny's Petit Mal [19] uses simple movements to engage people in rhythmic interaction. We focus on shadow puppet play as a simplified context for examining rhythm and nonverbal cues in bodily motion.

3. SHADOW PUPPET PLAY

According to the socially situated cognition perspective, communication is action oriented [5] and cognition is an adaptive process that is tightly coupled to action. Therefore, to participate in interactions, our robot must be able to process signals and adapt its models in real-time [4]. Taking a cue from human interaction, it makes sense to try and infer emotions and mental states from facial expressions and gestures. Unfortunately, since in embodied interaction there are many such channels of communication [5], it is not practical to capture and model all of these channels in real time. Manual transcription of recorded human data can be helpful in building models of interaction, however this process can be laborious and time consuming.

Our research draws on children's shadow puppet games, where the shadows of hands cast against a wall are used to express a story. Shadow puppets accommodate an open array of possibilities for interpretation and action while capturing the essential features of affect, meaning and intention that evoke human narrative propensities and emotional responses. Humans frequently perceive simple schematic artifacts as exhibiting a higher degree of sociality than their simple forms and actions contain [20]. Minimally designed robots that exhibit simple attentive, emotive and interactive gestures and spatial movements can convey the essential features of affect, meaning and intention in technology design [21]. As an interaction medium, shadow puppetry affords us the ability to observe an embodied discourse between two people that is expressive enough to support basic components of interaction and allows the participants to convey and infer the meaning of emotive gestures. At the same time, it limits the channels of communication to the point where we can hope to capture and model the signals in real-time using available computational and perception tools.

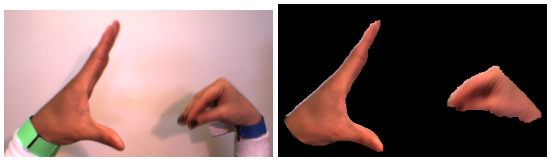
4. MODELING SHADOW PUPPET PLAY

Our aim in this project is to observe and model the behav-

ior patterns that characterize co-presence and coordination between interacting agents. For a robot, the ability to predict how a human will respond to its actions (and vice versa) is a form of coordination. Rather than using predictive frameworks that fully model the physical and mental states of the human to select actions, our robot uses predictive models that relate its actions to the behavioral responses of a human. We do not intend to model the high level cognition aspects of interaction and communication (such as the meaning of various gestures in context), but rather to automatically determine the parameters of the low level mechanisms (i.e. interaction synchrony, imitation, anticipation) that help to coordinate and ground interaction. The models described in [12, 2] highlight the difference between synchronization of content and synchronization of process and posit that the former is not possible without the latter. Our current work thus represents a foundation upon which more content-oriented models can be built.

In implementing shadow puppet play in a robot arm, we follow a three-part process which will be discussed in detail in the following sections. Firstly, we identify gestural primitives that comprise shadow puppet interactions by observing humans engaged in shadow play and develop automatic recognition capability for the relevant gestures. We use an online survey to evaluate the ability of third-person observers to classify interactive and non-interactive video segments. Second, we decompose the video sequences into streams of gesture tokens which correspond to the previously identified gestural primitives and measure occurrence and co-occurrences of behaviors. Finally, we use the data to build models of interaction for a shadow puppet robot.

4.1 Interaction data



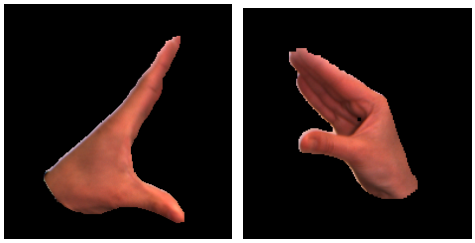
(a) Step 1 (b) Step 2

Figure 1: Steps 1 and 2 in the process of generating sample interaction sequences. We record interactions and remove all but the players' hands.

We start with the premise that humans are proficient at recognizing natural interaction. Using this idea, we attempt to model the behavior patterns in interaction sequences and define properties of these models that correlate with human evaluations. First, we video record examples of shadow puppetry between two people(figure 1a). We record the interaction in two separate video channels using a stereo camera[1]. This ensures that we have two time-synchronized videos where one of the players is centered in each of the channels. During the recording, each player wears a wrist marker, and each video frame is post-processed by segmenting out the wrist and hand of each player and removing all other data (figure 1b). The end result is two separate video sequences, one with only the left player and one with only the right player.

[1]Stereo is used for recording the players in separate channels, not for depth information

These processed video sequences represent our control data. Our experimental data is constructed by randomly recombining left and right sides of these sequences as in figure 2. The randomly stitched video sequences represent the experimental data. All videos consist of just the filled, moving outline of the players' hands and contain no features that can be used to visually distinguish between the two classes. Each video is approximately 25 seconds in length.



(a) Left (b) Right

Figure 2: Step 3: The players' motions are separated into individual video streams. This artificial sequence is generated by replacing the right player from the sequence in figure 1 with a player from another sequence.

Our next step is to collect evaluative feedback, which enables us to categorize our videos according to the observer's perception of interactivity. The use of outside observers to quantify interaction has precedence in experimental psychology([22], for example). We have designed a website that allows people to watch our processed interaction videos and rate the interaction (see [23]). The survey is modeled as a game, wherein the goal is to correctly identify the class of the video. There are a total of 24 video sequences, 12 of which are real interactions and 12 of which are artificial interactions. Players are asked to watch a sequence of 10 unique videos. The 10 videos are drawn uniformly at random from the set of 24, and the order of the videos presented to players is randomized. After each video, the players are asked whether or not the two participants could see each other during the interactions. Upon completing the survey, the participants are told how many videos were labeled correctly (but not which) and are invited to play again. For each video, the individual votes are tracked and turned into an interaction score by dividing the number of positive votes by the total number of times the video was rated. There were 382 total ratings, 284 correct, and 98 incorrect.

4.2 Gestural vocabulary

Our robot must be able to recognize and distinguish between significant patterns in the behaviors of the human participant in real time. We isolate the gestural primitives being used by observing, coding, and analyzing humans playing shadow puppets. We identify a gesture vocabulary composed of basic behavioral competences: nod, shake, talk, jerk, flick, touch. These gesture vocabularies are used in constructing the robot's perception of certain human gestures as well as in the robot's implementation of gesture.

We have developed a perception system that recognizes the basic motions used in the shadow puppet game. We use a simple colored wrist marker which allows us to automatically infer the wrist position of a player, and from that to infer the locations of the hand center and finger tips. The

hand contour is determined by searching the image for the skin colored blob that is nearest to the wrist marker. Once the hand contour is segmented, template matching is performed by comparing the extracted contour to a set of user specific predefined templates, by computing and comparing the Hu moment invariants for each contour [24]. If the hand is closed, the finger tips are located by finding the principle axes of the hand contour. If the hand is opened, the finger tips are located by finding the largest concavity of the contour. The wrist $w(t)$, hand center $h(t)$, and finger tip locations $f1(t), f2(t)$ as shown in figure 3 provide a rough kinematic model of the hand.

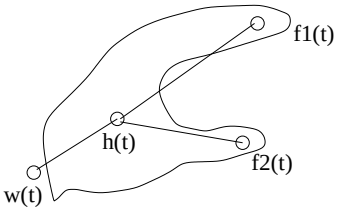


Figure 3: Hand model parameters

To automate the process of gesture recognition, we record the parameters of the kinematic model in each frame. At each instant of time, we parameterize the behavior of the human by measuring variations in the kinematic model over a recent history. The kinematic model of the hand at time t, can be described (non-compactly) by the state vector

$$[w(t), h(t), f1(t), f2(t)]^T$$

. The parameters used to identify behaviors are computed using a history of length $n = 7$ and measuring statistical dispersion of these parameters. The measures are given in equations 1– 4 , which are used to compute the 4×1 vector $b(t)$ of behavior parameters at each instant of time t (equation 5):

$$q1(t) = Var(w(i) \angle f1(i))_{i=t-n:t} \quad (1)$$

$$q2(t) = Var(\|h(i) - f1(i)\|)_{i=t-n:t} \quad (2)$$

$$q3(t) = Var(f1(i))_{i=t-n:t} \quad (3)$$

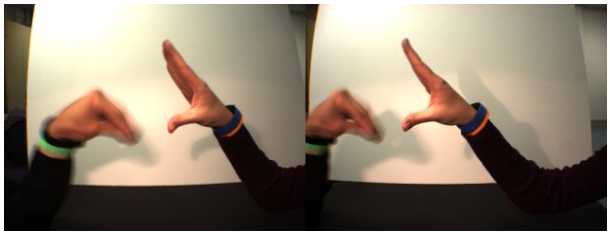
$$q4(t) = Cov(f1(i), f2(i))_{i=t-n:t} \quad (4)$$

$$b(t) = [q1(t), q2(t), q3(t), q4(t)]^T \quad (5)$$

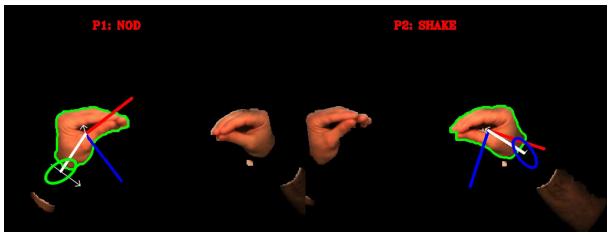
The behavior of the player is classified as either **Nod**, **Shake**, **Talk**, **Jerk**, **Flick**, **Touch**, or **None**. We calibrate our gesture recognition system for each user. During the training phase, the user performs each of the gestures several times. The vector of behavior parameters is computed and recorded for each of the examples.

4.3 Models of Interaction

We will now describe a method for modeling the low-level signal exchange in interaction using simple predictive models. The goal is to build models that can be used to generate behaviors that are interactive in the sense of being coordinated with a human partner. We explicitly restrict ourselves to modeling the exchange of signals, without imposing meaning on the signals. Figure 4a shows the experimental setup used to collect data. The participants wear wrist markers. Figure 4b shows the processed frames of the perception system. This system processes video and outputs the stream of behaviors in real-time.



(a) Dyadic Interaction



(b) Gesture Labeling

Figure 4: Experimental setup.

Using the behavior recognition system, the behavior of each player is converted to a one dimensional signal. Let

$$\Sigma = \{Nod, Shake, Talk, Jerk, Flick, Touch, None\}$$

be the set of possible symbols and let $X, Y \in \Sigma^*$ denote the behavior sequence of players 1, and 2 respectively. Let x_i and y_j denote the realizations of X and Y as behaviors i and j respectively. To model the interaction, we start by constructing the joint and marginal probability distributions of the player behaviors. In this case, the $P(x_i, y_j)$ represents the normalized frequency of two behaviors i and j, occurring at the same time, and $P(x_i)$ and $P(y_j)$ are the marginal probabilities of the behaviors occurring, independent of the other player's behavior. A 2D histogram is a convenient way to represent these joint and marginal probability distributions. $P(x_i, y_j)$ is the value of an individual histogram bin, with rows i, and columns j, normalized by the total number of samples. $P(x_i)$ is the normalized sum of a row, i and $P(y_j)$ the normalized sum of a column, j.

Recall that our end goal is to uncover measures of the interaction sequence that correlate well with the interaction scores assigned by the human observers. The distribution properties we have measured are conditional entropy (CE), Kullback-Leibler divergence (KL), and mutual information (MI), where

$$CE \equiv \sum_{\forall i,j \in \Sigma} Pr(y_j, x_i) log Pr(y_j | x_i) \quad (6)$$

$$KL \equiv \sum_{\forall i \in \Sigma} Pr(x_i) log(\frac{Pr(x_i)}{Pr(y_i)}) \quad (7)$$

$$MI \equiv \sum_{\forall i,j \in \Sigma} Pr(x_i, y_j) log(\frac{Pr(x_i, y_j)}{Pr(x_i)P(y_j)}) \quad (8)$$

For two random variables X and Y, the conditional entropy of Y on X represents the uncertainty (entropy) of the variable Y, when the value of variable X is given, averaged over all possible values of X. It is important to note that this is different than calculating the entropy of Y when X takes on a particular value, x_i. Kullback-Leibler divergence (also called relative entropy) is a measure of similarity between two dis-

MI	KL	CE $(X\|Y)$	CE $(Y\|X)$
0.5607	-0.5603	0.0751	0.0220

Table 1: This table shows the correlation of each distributions measure with the human evaluation of interactivity.

tributions. Mutual information measures the independence of two random variables, and is closely related to KL divergence and conditional entropy. Mutual information describes how the realization of a particular variable X reduces the entropy of another random variable Y. Mutual information does this by measuring the KL divergence between the joint distribution of the two variables and the product of their marginal distributions. This means that if the two random variables are conditionally independent, their mutual information is zero. Information theoretic properties have been used for comparing protein sequences in [25] and in [26] where mutual information kernels are applied to mixture models.

Figure 5 shows the user rating of each video sequence compared to the mutual information and KL divergence distribution measures. Each red cross represents a real interaction sequence, and each blue circle represents an artificial interaction sequences. The real sequences are numbered 1 through 12 and the artificial sequences are numbered 13 through 24. In each plot, the survey score is represented on the Y-axis. Table 1 shows the correlation of each measure with the survey score. Mutual information is strongly correlated with survey score, and KL divergence has a strong negative correlation with survey score. Correlation between conditional entropy and survey score is not significant. Based on the results from figures 5a and 5b we posit that interactive behavior is strongly correlated to high mutual information. We validate this theory using a human-robot interaction study using the Barrett Robot hand and Whole Arm Manipulator (WAM).

5. INTERACTIVE MODEL VALIDATION

Figure 6: Embodied interaction study: Subjects play the shadow puppet game with our 4 DOF Whole Arm Manipulator (WAM). A stereo camera is used to code the gestural language tokens of the human in real-time.

5.1 Control strategies

We next turn to the task of generating interactive control strategies from the existing behavior models. In the previously described form, behaviors X and Y of the players are modeled as random variables. The joint probability models can be used to answer queries about the probability of all possible random events involving X and Y. Our approach is to have our agent take on the role of player 2, and use the joint distributions to select actions which, according to Bayes rule, are most likely to occur during interactions between two human players. Given the behavior x_i of player 1, the agent assigns a value to Y by sampling from the distribution $P(Y|X = x_i)$.

In this study, the Barrett Robot hand and Whole Arm Manipulator (WAM) replace player 2 in the shadow puppetry game (figure 6,[27]). The WAM is a 4-degree of freedom system with human-like kinematics. The WAM is instrumented with a set of predefined gestural primitives that match those of the human player. Each behavior is executed by following a predefined trajectory that connects several points in the configuration space of the robot

In our interaction study, the human is asked to participate in 4 successive sessions of interaction with the robot. In each of the four sessions, the robot uses a different control strategy. The trials last for two minutes and in each trial, the order of the controllers is randomized. There are a total of $N = 8$ human subjects. The first controller (**C1**) samples from the distribution of an observed human-human interaction sequence, which has high user rating and high mutual information. The second controller (**C2**) samples from the distribution of an observed human-human interaction sequence, which has low user rating and high mutual information. The third controller (**C3**) simply imitates the human. A model which defines imitation, by definition, has highest possible mutual information, as well as KL divergence of zero. This strategy is selected to better understand the role of mutual information in interactivity. The fourth controller tested (**C4**) is a first order Markov model. To build a zero order Markov controller, we construct a Markov chain, in which the state of the system is defined by the pairwise behavior of the two players. Recall that behaviors are drawn from the alphabet

$$\Sigma = \{Nod, Shake, Talk, Jerk, Flick, Touch, None\}$$

. The first order model has $|\Sigma|^2$ states, and a $|\Sigma|^2 \times |\Sigma|^2$ transition matrix. The entries in the transition matrix represent

$$Pr(X_t = x_j, Y_t = y_j | X_{t-1} = x_{t-1}, Y_{t-1} = y_{t-1}), \forall x_i, y_j \in \Sigma$$

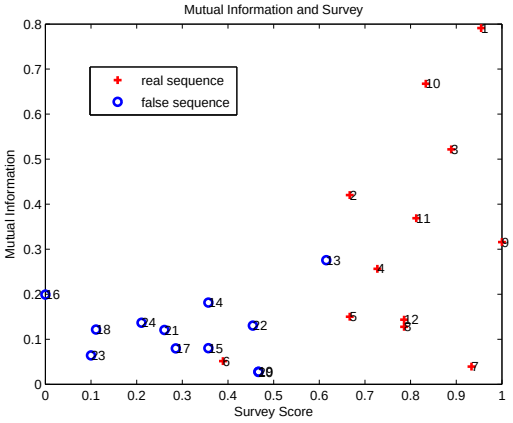
the probability of observing an event $X_t = x_j, Y_t = y_j$, given some previous observation $X_{t-1} = x_{t-1}, Y_{t-i} = y_{t-1}$. This is in contrast to the previous reactive models, which define only

$$Pr(X_t = x_j, Y_t = y_j), \forall x_i, y_j \in \Sigma$$

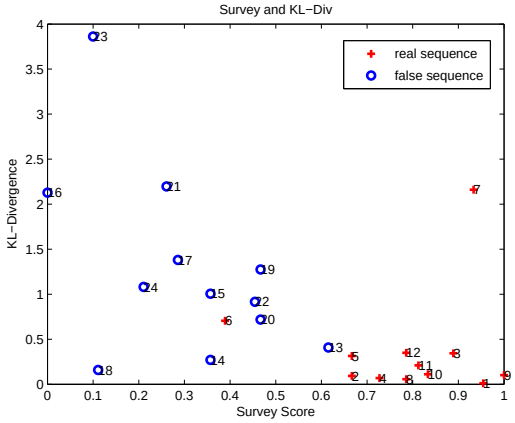
the probability of observing an event occurring independent of previous events.

5.2 Results

Each person interacting with the robot is asked to rate a set of three statements after each of the four trials: **(1)** *The robot reacted appropriately to me*; **(2)** *The robot could recognize my actions*; **(3)** *The robot seemed intensely involved in the interaction.* The statements are aimed at gauging the user's perception of the robot's ability to recognize gestural cues and react accordingly, as well as its social presence in the interaction. The subjects rate their agreement with each question by assigning a value between 1 and 6. The average rating of each controller, for each question is displayed in figure 7.



(a) Interaction score vs. mutual information



(b) Interaction score vs. Kullback-Leibler divergence

Figure 5: Figure(5a): Scatter plot of interaction score and mutual information for each of the 24 video sequences. Figure(5b): Scatter plot of interaction score and KL divergence for each of the 24 video sequences.

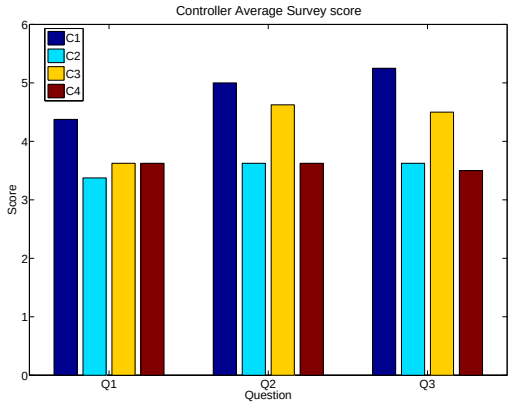


Figure 7: Average response to each question for each of the four controllers ($N = 8$).

The results indicate that the high mutual information controller (C1) generates the most interactive behavior. Surprisingly, subjects judged the recognition rate of controller (C1) to be on average as good as the recognition rate of the imitation controller (C3). The scores for the controller with high mutual information and low survey score (C2) and the first order Markov model (C4) are both significantly lower than controllers C1 and C3.

We have also measured the mutual information and KL divergence of the behavior sequence data in our interaction study. We calculate this in the same way that it is calculated in the human-human interactions. In each trial, the gestures of the human and robot are recorded and a histogram of co-occurrences is computed. Figures 8a and 8b show the measured mutual information and KL divergence of each test subject in each sequence. The X-axis represents the subject number, and each of the four lines represents one of the four controllers used.

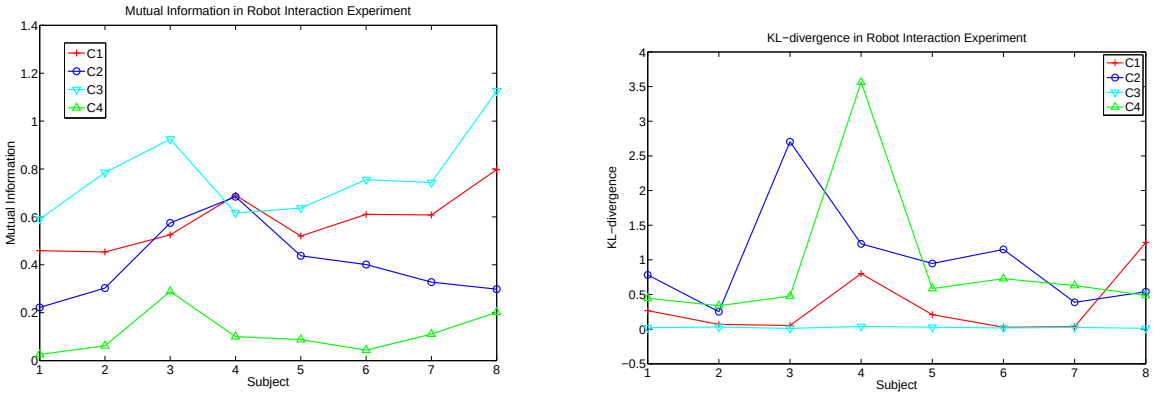
In all but one case (subject 4), the mutual information for the imitation controller (C3) is higher than for the other 3 controllers. This is consistent with the survey score and what we expect to see from a controller that imitates, since the behavior of one player should be perfectly informative about the behavior of another player.

For the learned controller with high mutual information and high survey score (C1), the measured mutual information is second highest after C3 in all but two cases. For subject 4, C1 has highest mutual information, tied with C2. For subject 3, C2 generates slightly higher mutual information than C1. In all other cases the C2 is lower than C1 and higher than C4. This is consistent with the measured mutual information from the human-human interaction study, as is the relatively low survey score.

Recall that KL divergence is measure of the similarity between the marginal distribution of the two players' behaviors. As one might anticipate, the measured KL divergence for the imitation controller (C3) is close to zero in all cases. The high mutual information controller (C1) is also low in all cases, except for subjects 4 and 8. This is interesting, because for those two subjects, the measured mutual information is highest. Also, note that the measured KL divergence of C2 tends to be significantly higher than that of C1 for all cases except for subject 8.

5.3 Discussion

For the first-order Markov controller, mutual information is lower than other controllers for all test subjects. This is consistent with the ratings given by the users. In general, it would appear that this is the worst of the four controllers by all measures. However, it should be noted that the process used by the first-order Markov model to generate behaviors is considerably more complicated than the other three controllers. We have observed several cases, where the human tries to figure out the pattern of the robot's behavior, and often succeeds in doing so. For example, when the human is making a gesture, the response of the robot is often to do nothing, wait until the human stops, and then respond. This is a form of learned turn taking that is implicit in our model and control methods, and it occurs frequently when using the first-order Markov model controller. In several cases the



(a) Mutual Information of joint distribution of human and robot behaviors

(b) KL-Divergence of human and robot behavior distributions

Figure 8: Interactivity Measures in Human-Robot interaction behavior sequences.

human repeats a behavior, apparently trying to evoke a response. Often they figure out that they must pause and give a turn, after which the interaction is more fluid. It may be the case that more interaction time is required for the human to understand the pattern of behaviors generated by this more complex model.

Our interaction models are built by observing the co-occurrence of gestural tokens. The implicit assumption is that interaction should be modeled by the joint behavior of the the two players together at each instant of time. However, due to the temporal lag between actions and reactions, it may be more useful to model the joint behavior of one player at an instant of time and the behavior of the other player at some point after that. For example, rather than align the two behavior sequences so that we measure the relationship between X_t and Y_t we might shift the alignment by some length n, so that we measure the relationship between X_t and Y_{t+n}. To justify the use of a zero shift, we have done the following: For each example sequence, we compute the mutual information over a range of different alignments. We select a window of size n and for each $i \in -n : n$, we build a histogram where each bin $[row_i, col_j] : (i, j) \in \Sigma$ counts the occurrences where $X_t = x_i$ and $Y_{t+i} = y_j$. We measure the mutual information of the histogram at each iteration of i, and find the value of i for which mutual information is maximum. In doing this, we have observed that the value of mutual information at each such index tends to be normally distributed with mean close to zero. This justifies the use of zero shift.

While we do not address the content of interaction in this study, the synchronization of action and response that we describe serves as the foundation for more complex interaction and communication. In our initial observational study of human shadow puppet play, we saw that different rhythms and gesture combinations were involved in enacting various interaction schemas, which could be interpreted as the shadows fighting, playing, having a conversation. In the case of collaborative human-robot interaction scenarios, the ability of both partners to synchronize their actions serves as the foundation for joint attention, shared meaning, and contextual grounding. Interactors can miss or misunderstand cues when they are not presented in a mutually established flow of interaction. In this paper we describe how people attribute social presence and interactivity to the robot. Our next steps will be to extend our model to enable recognition of and participation in these different content-oriented interactive schemas (e.g. argument, conversing, etc.).

A socially interactive robot was used as a test-bed for theories and models of interactivity, human social behavior, and the attribution of human characteristics to non-humans. We now have a basic computational model of synchronous, contingent interaction and the situated dynamics of human social behavior. Much as infants have evolved to take advantage of their caregivers' knowledge and propensity to attribute various intentional and affective states, a robot that collaborates seamlessly with humans in everyday activities has to be able to take advantage of the human's knowledge of the world and adapt its behavior accordingly. New perception, decision-making and control algorithms based on generative models of joint interaction, such as the one described in this paper, should be designed to achieve this capability.

6. CONCLUSIONS

In this paper, we have studied the fundamental behavioral patterns and cues that enable the development of social attachment and collaborative interaction. We also contribute to the field of human-robot interaction, by developing a formal description of some of the fundamental aspects of interaction that would enable robots to perform as communicative partners. To achieve these cross-disciplinary goals, we develop computational models of synchrony in nonverbal interaction that simulate the underlying dynamics of humans social behavior and can be implemented in socially interactive robots.

We have evaluated methods for quantifying and generating interactive behavior for a robot. From the initial web survey, we can conclude that humans are proficient at identifying real interaction and that mutual information is a significantly better predictor of interactivity score than conditional entropy. There are several conclusions we can draw from this information. First, because the real sequences have high mutual information, the player behaviors are not independent and mutual information is useful for quantifying interaction. Second, in the real sequences, KL divergence tends to be low, so in real interaction the distributions of the player's behaviors are similar. Third, if we consider the subtle difference be-

tween mutual information and conditional entropy, we can conclude that it is not advantageous to rely on discriminative methods to model dyadic interaction. Specifically, measures of interactivity that rely on the behavior of one player X to directly predict the behavior of another player Y do not perform well. Instead we should build generative models of the interaction process as an intermediate step to answering queries like $Pr(Y|X = x_i)$.

Lastly, the results of our interaction study suggest that the interaction models and measures proposed in this work can be learned and utilized by a robot to generate interactive behavior. These models can help us design engaging socially interactive robots, as well as develop formal models of behavioral coordination that can be validated through embodied human-robot interaction. Currently, the control strategies used by the robot are responsive to the human, but not adaptive. In the future, we will be extending the robot's capabilities so that it will be able to not only perceive and automatically code what is going on during the interaction, but also continue populating the model with data constructed in interaction and to track changes in the model. Such a robot can be used as a controllable tool for further experimentation and in-depth study of the particular factors in social interaction, such as imitation, rhythmic entrainment, joint attention, and coordination.

7. REFERENCES

[1] C. L. Breazeal, *Designing Sociable Robots*. Cambridge, MA: MIT Press, 2002.

[2] G. R. Semin, "Grounding communication: Synchrony," *Social Psychology: Handbook of Basic Principles 2nd Edition*, pp. 630–649, 2007.

[3] K. Dautenhahn, B. Ogden, and T. Quick, "From embodied to socially embedded agents—Implications for interaction-aware robots," *Cognitive Systems Research*, vol. 3, pp. 397–428, 2002.

[4] L. W. Barsalou, C. Breazeal, and L. B. Smith, "Cognition as coordinated non-cognition," *Cognitive Processing*, vol. 8, no. 2, pp. 79–91, June 2007.

[5] E. R. Smith and G. R. Semin, "Socially situated cognition: Cognition in its social context," *Advances in Experimental Social Psychology*, vol. 36, pp. 53–117, 2004.

[6] S. Sabanovic, M. P. Michalowski, and L. R. Caporael, "Making friends: Building social robots through interdisciplinary collaboration," in *Multidisciplinary Collaboration for Socially Assistive Robotics: Papers from the 2007 AAAI Spring Symposium, Technical Report SS-07-07*. AAAI, March 2007, pp. 71–77.

[7] W. S. Condon and L. W. Sander, "Synchrony demonstrated between movements of the neonate and adult speech," *Child Development*, vol. 45, no. 2, pp. 456–462, 1974.

[8] D. McNeill, *Gesture & Thought*. Chicago: Chicago University Press, 2005.

[9] C. Crick, M. Munz, T. Nad, and B. Scassellati, "Robotic drumming: Synchronization in social tasks," in *IEEE ROMAN*, 2006, pp. 97–102.

[10] M. P. Michalowski, S. Sabanovic, and H. Kozima, "A dancing robot for rhythmic social interaction," in *Human-Robot Interaction Conference*, 2007.

[11] H. Ogawa and T. Watanabe, "Interrobot: speech-driven embodied interaction robot," *Advanced Robotics*, vol. 15, pp. 371–377, 2001.

[12] H. Clark and S. A. Brennan, *Perspectives on socially shared cognition*. Academic Press, 1991, ch. Grounding in communication.

[13] L. Barrett, "Too much monkey business," in *Proc. Grounding Sociality: Neurons, Minds, and Culture*, 2007.

[14] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge: Cambridge University Press, 1995.

[15] L. R. Caporael, "The evolution of truly social cognition: The core configuration model," *Personality and Social Psychology Review*, vol. 1, pp. 276–298, 1997.

[16] E. Z. Tronick and J. F. Cohn, "Infant-mother face-to-face interaction: age and gender differences in coordination and the occurrence of miscoordination," *Child Development*, vol. 60, no. 1, pp. 85–92, 1989.

[17] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, pp. 893–910, 1999.

[18] P. Andry, P. Gaussier, S. Moga, J. P. Banquet, and J. Nadel, "Learning and communication via imitation: An autonomous robot perspective," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 31, no. 5, pp. 431–442, 2000.

[19] S. Penny, "Embodied cultural agents: at the intersection of robotics, cognitive science, and interactive art," in *AAAI Socially Intelligent Agents Symposium*, 1997.

[20] M. Blow, K. Dautenhahn, A. Appleby, C. L. Nehaniv, and D. Lee, "The art of designing robot faces: Dimensions for human-robot interaction," in *HRI '06: Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. New York, NY, USA: ACM, 2006, pp. 331–332.

[21] H. Kozima and C. Nakagawa, "Social robots for children: Practice in communication-care," in *Proc. IEEE Intl. Workshop on Advanced Motion Control*, 2006, pp. 768–773.

[22] F. J. Bernieri, J. S. Reznick, and R. Rosenthal, "Synchrony, pseudosynchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions," *Journal of Personality and Social Psychology*, vol. 54, no. 2, pp. 243–253, 1988.

[23] Meisner and Sabanovic, "Shadow puppet interaction site," May 2008, www.cs.rpi.edu/~meisne/interaction.

[24] M.-K. Hu, "Visual pattern recognition by moment invariants," *Information Theory, IEEE Transactions on*, vol. 8, no. 2, pp. 179–187, 1962.

[25] M. Cuturi and J.-P. Vert, "A mutual information kernel for sequences," *Proc. IEEE Intl. Joint Conf. on Neural Networks*, vol. 3, pp. 1905–1910 vol.3, July 2004.

[26] M. Seeger, "Covariance kernels from bayesian generative models," in *NIPS*, vol. 14, 2002, pp. 905–912.

[27] Meisner, www.cs.rpi.edu/~meisne/interaction/wam.html.