Who is IT? Inferring Role and Intent from Agent Motion

Christopher Crick
Yale University
New Haven, CT
christopher.crick@yale.edu

Marek Doniec
Yale University
New Haven, CT
marek.doniec@yale.edu

Brian Scassellati
Yale University
New Haven, CT
scaz@cs.yale.edu

***Abstract*— We present a system which observes humans interacting in the real world and infers their goals and intentions through detecting and analyzing their spatiotemporal relationships. Given the motion trajectories of human agents and inanimate objects within a room, the system attempts to characterize how each agent moves in response to the locations of the others in the room – towards an object, say, and away from the other agent. Each of these calculations leads to an estimate of the agent's current intentional state. Taken together with the other agents in the room, and paying particular attention to the moments when the various agents' states change, the system can construct a coherent account of the action similar to the stories humans tell. We illustrate this approach with a robot that watches people playing a game of tag, works out for itself the roles and intentions of the various players, and then attempts to join in the fun. The robot's interpretation of events agrees with human observers on average 70.8% of the time – nearly as good as the agreement between two humans (78.5%).**

***Index Terms*— Intention recognition, motion, causation**

I. Introduction

Psychologists have long known that humans possess a well-developed faculty for recognizing dramatic situations and attributing roles and intentions to perceived characters, even when presented with extremely simple cues. A human will watch three animated boxes move around on a white background, and describe a scene involving tender lovers, brutal bullies, tense confrontations and hair-raising escapes. Furthermore, a wide variety of human observers will construct the same dramatic story out of the same ludicrously simple animation. Our ability to make sense of a scene does not seem to depend on rich contextual information, lending credence to the idea that a machine might be able to accomplish the same sort of inference.

Of course, a problem arises immediately in attempting to bring computational resources to bear upon this same data, if we wish to interpret activity and interaction in the real world rather than in artificially constructed animations and simulations. Visual perception is an enormous problem in its own right, and though this area is a subject of intense research, we still understand only pieces of the process. A human's visual system translates retinal impulses into coherent, persistent, reliable object concepts which can be used to reason about the world in terms of space and motion. At the current state of the art, however, computational perception cannot dependably accomplish the same task.

Fig. 1. The robot Nico.

Does this mean, then, that we cannot investigate the development of embodied, real-world cognition about complex social behaviors until we have completely solved the problems of perception which stand in the way? No, but it does suggest that we must endow our robots with other means of acquiring the information they need if we are to make any progress. Today's most advanced robotic navigation systems still require laser rangefinders and satellite triangulation to accomplish what human drivers can do with their eyes. Likewise, to bypass some of the complexities of visual perception, we use a sensor network to supplement the robot's cameras with regular position updates for the people and objects it watches.

The platform upon which we have implemented our developmental model is the robot Nico (see Figure 1). It is an upper-torso humanoid modelled after the body dimensions of a one-year-old child, possessing seven degrees of freedom in its head and six in in each arm. Nico has been used for a wide variety of research into childhood development of kinematics [1], language acquisition [2] and theory of mind [3], as well as investigating social interaction among humans such as the development of gaze following [4].

Nico's cameras give it stereoscopic vision, both a wide-angled peripheral and a highly-detailed foveal image. From the perspective of modelling the development of human

social cognition, of course it would be best to sense the whereabouts of the people it observes by using this visual input exclusively. The precise positioning data provided by our sensor network will hopefully serve as a baseline for bootstrapping the more complex visual perceptual problem.

II. Related Work

The original inspiration for the approach we take comes from the pioneering psychology study by Heider and Simmel more than a half-century ago [5]. They found that humans, given very rudimentary visual cues in the form of simple animations, would happily spin complex tales about the events they witnessed. Furthermore, these stories were remarkably consistent from one subject to another. We are powerfully motivated to anthropomorphize the interactions we witness, and to attribute social roles and intentions to candidates as unlikely as boxes on a white screen.

Our work also draws from more recent literature in developmental psychology, investigating the fundamental cognitive processing modules underpinning perception and interpretation of motion. These modules appear responsible for our rapid and irresistable computation of physics-based causality [6] [7], as well as facile, subconscious individuation of objects in motion independently of any association with specific contextual features [8] [9] [10].

The specific analysis undertaken by our system, hypothesizing vectors of attraction and repulsion between agents and objects in the world in order to explain the causal relationships we note in an interaction, relates to the dynamics-based model of causal representation proposed by Wolff [11] and on Talmy's theory of force dynamics [12]. Humans can explain many events and interactions by invoking a folk-physics notion of force vectors acting upon objects and agents. This holds not only for obviously physical systems (we talk easily of how wind direction affects the motion of a sailboat), but for social interactions as well (the presence of a policeman can be interpreted as an actual force opposing our desire to jaywalk). Our system explicitly generates these systems of forces in order to make sense of the events it witnesses.

Gerd Gigerenzer and Peter Todd have used simple animations of the Heider and Simmel variety for testing simple social classification heuristics [13], [14]. They asked subjects to generate animations via a game: two people sat at computers and controlled two insect-like icons using a mouse, producing simple socially significant scenarios such as "pursuit and evasion", "play" or "courting and being courted". Using these animations, they generated features such as relative heading and absolute speed and used very simple decision mechanisms to identify the social context behind the animation. Their approach showed that indeed machines could recover this sort of social information from simple animations. However, they largely intended their experiments to demonstrate, successfully as it happens, the efficacy of their simple heuristics in relation to more computationally complex classification mechanisms. Their scenarios were very stylized, with a very narrow repertoire of possibilities, and while humans created the animations, they did so in an artificial machine-mediated fashion. Our approach, on the other hand, provides us with a record of real human activity in the world. Our scenarios are richer, and the level of detail is much greater. Rather than simple classification among a half-dozen intentions, our system generates a genuine narrative.

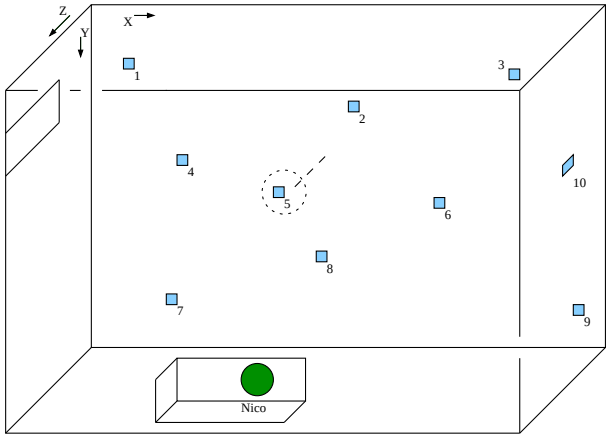

Fig. 2. Schematic representation of lab layout, from above. Sensors 1-4 and 6-9 are affixed to the ceiling ($z = 0$). Sensor 5 hangs from an overhead light, 95 cm below the ceiling, and sensor 10 is attached to a wall at $z = 128$ cm.

III. Setup

Most of the literature examining how intentions and goals can be inferred from motion cues relies on canned or simulated worlds. From Heider and Simmel's manually-constructed animations to Peter Todd's computer bug game, researchers have studied humans' and computers' abilities to classify and interpret simple, abstract animations by generating such animations in simple, abstract ways. In contrast, our data derive from real-world human interaction, obtained with a localizing sensor network of radio- and acoustic-enabled Cricket nodes.[1] These devices send messages to one another using a simultaneous radio broadcast and ultrasound chirp, and the receiving unit can calculate distance by comparing the difference in arrival times between the two signals. With a beacon node corresponding to each agent involved in a particular game or dramatic scenario, and ten listener nodes in fixed positions throughout the room, we can determine a person or object's location within a few centimeters through triangulation. Each participant carried or wore a uniquely-identified sensor throughout a particular scenario, and the toys or objects they manipulated were likewise tagged.

[1]See [15] for hardware details. We entirely redesigned the embedded control systems and data analysis software to meet the needs of our experiments.

This sensor system produces a five-dimensional vector of data: the identity of an object, its x, y and z coordinates in the room, and the time at which these coordinates were measured (accurate to the millisecond). Figure 2 shows the coordinate frame and the robot's position with respect to the rest of the room. The z coordinate, measuring vertical distance from the lab's ceiling, is only used for appropriate distance measuring and robot-human interaction; a two-dimensional top-down representation is perfectly adequate to interpret the activity. Our robot uses this data to perform its analysis of the interactions it sees, and to respond appropriately.

In addition, we use the data to construct simple Heider-and-Simmel-style animations representing the scenario. Our system generates a colored square for each agent and interpolates between detected positions to create an animation for human evaluation. From a record of the raw position reports coming from sensors, we generate a Scalable Vector Graphics (SVG) file specifying an animation which can then be viewed in a web browser. A test subject can view the video and click with the mouse to specify her interpretation of the events she sees, with the system capturing the times and locations of the clicks. In this way, we can directly compare the human and robot ability to generate hypotheses concerning the witnessed activity, using exactly the same perceptual data. In most cases, the interpolation of the motion between the detected positions is simply linear, but for some of our test runs, we used cubic splines to generate a smoother cartoon. Our human test subjects had no trouble interpreting the simpler linear model, however, though most test subjects reported that the smoother animation looked subjectively better.

To be considered as an appropriate model of social cognition, our system should produce the same analyses from the same data as the subjective impressions of the humans watching these cartoons. Provided that humans can readily interpret the intentions, goals and dramatic moments in animated representations of simple four-dimensional data, our machine-based algorithm should be able to do the same. To do this, we take our cue from psychology and folk physics and cause the robot to imagine "forces" acting on the people it sees, causing them to act the way they do.

Our basic approach involves generating a set of hypotheses for each agent's attractions and antipathies toward the other objects and agents in the room, and recognizing the points at which these hypotheses needed to change in order to continue to represent the data well. We assumed that each agent moved in accordance with these qualities – in other words, at each moment in time, an agent's velocity (or, alternately, acceleration) is the sum of the attractive and repulsive vectors associated with each of the other agents and objects in the scenario. The algorithm searches for the best fit in a least-squares sense among competing hypotheses over a short time frame – 2-5 seconds. Additionally, the influences between agents could be inverse, inverse-square or constant.

To make these calculations, we proceed as follows:

1) Calculate the current velocity of an agent separately for each dimension: $V_{x_i^n} = \frac{x_i^{n+1} - x_i^n}{t_{n+1} - t_n}$ (and similarly along the y dimension). Here, $V_{x_i^n}$ represents the x component of agent i's velocity at time n. x_i^n, x_j^n and x_k^n are the x coordinates of agents i, j and k respectively, at time n. Below, d_{ij}^n and d_{ik}^n are the Euclidean distances between i and j or i and k at time n. The extra d in the denominators comes about because simple coordinate subtraction introduces distance into the numerator, and must be normalized out.
2) Assume that the agent's velocity is determined by a combination of influences from the other participants in the scenario, represented by one of a small number of equations:
   - Constant: $V_{x_i^n} = \frac{c_{x_j}(x_j^n - x_i^n)}{d_{ij}^n} + \frac{c_{x_k}(x_k^n - x_i^n)}{d_{ik}^n} + \cdots$ (and similarly along the y dimension)
   - Inverse: $V_{x_i^n} = \frac{c_{x_j}(x_j^n - x_i^n)}{(d_{ij}^n)^2} + \frac{c_{x_k}(x_k^n - x_i^n)}{(d_{ik}^n)^2} + \cdots$ (and, again, similarly along y)
   - Inverse squared: $V_{x_i^n} = \frac{c_{x_j}(x_j^n - x_i^n)}{(d_{ij}^n)^3} + \frac{c_{x_k}(x_k^n - x_i^n)}{(d_{ik}^n)^3} + \cdots$ (same with y)
3) Collect all of the data points falling within a short time period into a pair of matrices. For three total agents involved, using the constant-influence hypothesis, and with a window of $\delta n = 2$ seconds, these matrices would be:

$$A = \begin{vmatrix} \frac{x_j^n - x_i^n}{d_{ij}^n} & \frac{x_k^n - x_i^n}{d_{ik}^n} \\ \frac{y_j^n - y_i^n}{d_{ij}^n} & \frac{y_k^n - y_i^n}{d_{ik}^n} \\ \frac{x_j^{n+1} - x_i^{n+1}}{d_{ij}^{n+1}} & \frac{x_k^{n+1} - x_i^{n+1}}{d_{ik}^{n+1}} \\ \frac{y_j^{n+1} - y_i^{n+1}}{d_{ij}^{n+1}} & \frac{y_k^{n+1} - y_i^{n+1}}{d_{ik}^{n+1}} \\ \frac{x_j^{n+2} - x_i^{n+2}}{d_{ij}^{n+2}} & \frac{x_k^{n+2} - x_i^{n+1}}{d_{ik}^{n+1}} \\ \frac{y_j^{n+2} - y_i^{n+2}}{d_{ij}^{n+2}} & \frac{y_k^{n+2} - y_i^{n+2}}{d_{ik}^{n+2}} \end{vmatrix}, b = \begin{vmatrix} V_{x_i^n} \\ V_{y_i^n} \\ V_{x_i^{n+1}} \\ V_{y_i^{n+1}} \\ V_{x_i^{n+2}} \\ V_{y_i^{n+2}} \end{vmatrix}$$

   The matrices representing the other two hypothesis domains are constructed similarly.
4) Construct a QR-decomposition of matrix A, and use it to find the least-squares solution of $A \times X = b$. The matrix X thus found corresponds to the best-fit constants for the hypothesized equations of motion.

The constants found by this process reflect the strengths of the relative force vectors in play at a particular point in time as they influence the motion of the agents in the scene – not necessarily physical forces, of course, but rather the conceptual forces arising from the social situation and the agents' goals. These constants form a set of features which the robot uses to segment and identify the events it watches. Moments when the constants change sign, or vary

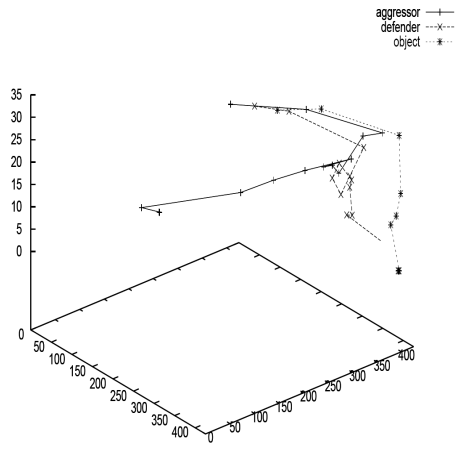

Fig. 3. A scenario where the defender tries to prevent the aggressor from snatching the object, but ultimately fails. The object remains still while the aggressor approaches and the defender moves to interpose itself. The aggressor obtains the object and moves rapidly away, followed by the defender.

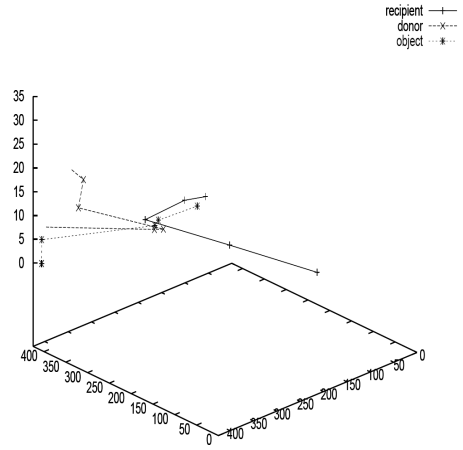

Fig. 4. A scenario where the donor hands an object to the recipient. The object accompanies the donor as the two agents approach each other, and then leaves with the recipient as they walk away.

dramatically in magnitude, or where the set of equations that produce the best data fit changes, are particularly important, and often corresponding to significant events in the scenario. These are usually the same moments that human observers mark, as well, as the experiments described below bear out.

This sort of computation is not scalable to arbitrarily large numbers of participants, though the matrices grow in a fairly reasonable $O(n^2)$ fashion. On the other hand, it does not seem plausible that humans are able to interpret the individual roles of every participant in a crowd scene, either. Rather, they tend to analyze large-scale interactions in terms of groups of people with similar roles and goals, and our approach can be extended to accommodate the same kind of evaluation.

Not only is the robot able to detect and interpret the events it sees, but it can respond and participate as well. We implemented a repertoire of gestures and behaviors that the robot can use in real time while observing a social interaction. The robot's motor system acts on information from the sensor tracking system and the intention recognition system. While the robot constantly receives sensor position data from the tracking system, the intention recognition system can induce Nico to perform the following behaviors:

- **Visual Tracking:** Nico can visually track a particular agent or object by turning its head to follow a sensor's movement about the room. At any time, the system can substitute one object of attention for another, or leave off visual tracking entirely.
- **Arm Tracking:** Nico stretches its right arm to point at a particular object or agent. This behavior is controlled flexibly and dynamically, according to the robot's current evaluation of the events and intentions it perceives.
- **Avoidance Behavior:** Nico can decide to show avoidance behavior towards the sensor attached to one of the objects or people in the room. It will look away from whatever it is avoiding and shield its face with an arm. This behavior supersedes the previous behaviors.

The first two behaviors can be combined – Nico can watch and point at different things simultaneously – but the avoidance behavior can only be executed exclusively.

We measured the absolute positions of the head (H) and the right shoulder (S) in the coordinate system that the cricket sensors use. To track an object with position X either visually or with the arm we geometrically compute the difference vectors $D_H = X - H$ or $D_S = X - S$ respectively. Using the difference vectors and a geometrical model of the robot we compute the desired motor positions. The head always uses two degrees of freedom (pitch and yaw). For the arm we usually use only two degrees of freedom in the shoulder (move arm up/down, left/right). However if the sensor we point to is far to the right the shoulder joints reach it's limits and we use two elbow joints in addition to point towards the target sensor with the lower arm.

IV. Results

A. Experiment 1

Our first experiment explored the feasibility of the force-dynamic approach in qualitative fashion. We devised simple scenarios such as "Person A wants the toy, but Person B tries to prevent him from getting it," and asked subjects to act out the events described. We collected the motion trajectory data and generated both animations and vector hypotheses.

We first asked human observers to watch the cartoon and describe what happens in their own words. Although the descriptions were somewhat idiosyncratic, most identified the same relationships between the pictured entities and noticed the same dramatic moments where those relationships

change. For instance, for the interaction illustrated in Figure 3, one observer commented "The red guy wants to get the blue thing. First he just walks up, but then the green guy tries to get in the way. They dance around a bit, and then the red guy gets past, grabs the blue square and takes off. The green one chases him." Different subjects used slightly different vocabularies, but told the same story and noticed the same events. The green square "got between" or "fought off" or "tried to stop" the red one, while the red square, 28 seconds into the animation, always succeeded in "taking" or "stealing" the blue object.

We allowed Nico to analyze the same data and produce its estimates of the time-varying constants and equations of motion which describe the agents' relationships with each other. We provided basic English phrases for the robot to use to describe its conclusions, selected according to the values of the mathematically-determined constants. At every significant event (whenever the hypothesized vectors changed sign or a different set of equations produced a better fit), the robot produced a new phrase. For the scenario just described, Nico was able to produce the following: "Red approaches blue. Green approaches red and approaches blue. Red avoids green and approaches blue. Green approaches red and avoids blue. Red and blue avoid green. Green approaches red and approaches blue." After the human subjects produced their own scenario descriptions, they were asked to answer the question "Was Nico's summary accurate (yes or no)?" They responded positively 2/3 of the time.

In order to obtain a more objective measure of how clear and interpretable our animations are, we provided a set of descriptions of the scenarios depicted in animations, as well as a few descriptions that did not correspond to any animation. For example, we described the scenario shown in Figure 4 as "Red gives green the blue toy." We asked our subjects to match descriptions and animations. All six subjects chose exactly the same description for each of ten animated scenarios, validating our assumption that humans are quite adept at understanding complex social interactions with very impoverished context and sensory data, and moreover different people's interpretations of the same data are largely consistent with one another.

B. Experiment 2

We devised our second experiment to take a closer, more quantitative look at the robot's performance in a longer, more complex motion scenario, and to give the robot the opportunity to participate meaningfully in the action. To that end, we chose to focus on watching and learning about the game of Tag. As Nico is currently not ambulatory, its participation in the game must be somewhat stylized, but it can still take appropriate actions that are readily interpreted as relevant to the game by human observers. While a game is being played, Nico looks at the person it currently judges to be "IT", while pointing to another player, the one farthest away from the robot itself – as if to say, "Hey, you, go over there. Get that guy, not me!" If the person Nico perceives as "IT" comes within two meters of itself, it assumes a defensive posture, throwing its arm over its face and looking away.

Figure 5 shows the robot's performance over the course of a 45-second game. We also converted this game and others to an animation and showed them to three human subjects, who were instructed to indicate the identity of "IT" as participants tagged each other back and forth over the course of the game. One of these sequences is also shown in the figure for comparison. Compared across four games and three subjects, the robot and humans agreed 70.8% of the time. This compares with an agreement between the human coders alone of 78.5% – that is, when watching a cartoon of a typical 45-second game of tag, two humans will disagree with each other about the identity of "IT" for a half a second here, a couple seconds there, for a total of about 9.8 seconds of incongruity. Both the human and robot performances are dramatically above random (33%).

The small disparity in performance likely arises from the fact that Nico has no preconceived notions about how the game of Tag operates. Specifically, it cannot deduce that "IT" can only pass between players when they are next to each other, unless such a pattern emerges from the games it sees. Human observers know this *a priori*. Such a discrepancy is illustrated in Figure 6.[2] At $t = 9$, blue (which had been chased by red) now moves toward green, while green runs away. Both human and robot judge that red has tagged blue and blue is now "IT". At $t = 13$, however, blue is still chasing green a little, but red begins moving quickly toward the other two. The computer assumes this means red is now "IT", but the human observer knows that red was too far away to have been tagged. Shortly afterwards, red and blue start running away from green, and it is obvious to both human and robot that a new player has become "IT".

V. Conclusion

Humans have a powerful capacity to construct rich, consistent interpretations of very impoverished data, so long as they have motion cues to work from. We have shown that, provided we can get around the difficulties in perceiving motion in the first place, machines are able to draw many of the same conclusions. We have demonstrated a robotic system that can infer roles, intentions and events based on hypothetical social "forces" constructed in real time from real-world motion data. The verisimilitude of the data thus collected enables us to draw stronger conclusions with respect to real human interaction and interpretation, in contrast to data derived from simulation or computer-mediated play.

[2]The colored squares are exactly what the human coders see in the animations. The arrows have been added for clarity on the page.

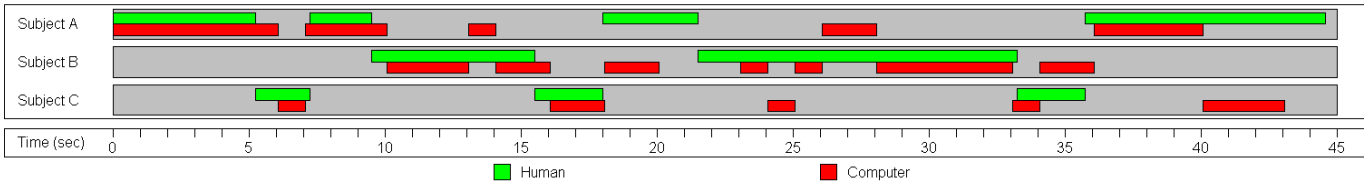

Fig. 5. Robot's opinion of who is "IT" in a game of tag, compared to a single human's analysis of identical data. The agreement between human and robot is closest in this sample to the statistical average, out of twelve total (three human coders × four games).

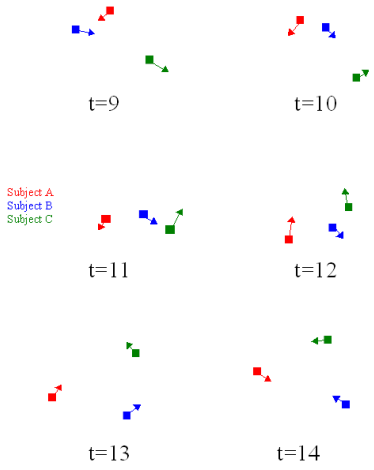

Fig. 6. Stills excerpted from game represented in Figure 5. See text for details.

We have so far investigated only simple scenarios with a fairly limited range of activity, but we have nevertheless demonstrated how much information can be gleaned from remarkably simple input. Furthermore, we have implemented a computational mechanism which draws conclusions that reasonably parallel our own human judgments. This approach can be extended in a number of ways. We would like to investigate how a system could learn a larger vocabulary of roles and intentions by observing and reasoning about its own inferences. For example, after watching enough Tag, the robot may be able to work out for itself that touch is an important component of the game, and begin looking for instances of touching on its own. We would also like to develop better algorithms for extracting this sort of motion from vision, using location information coming from the sensor network as scaffolding. Armed with conjectures about the roles and intentions of the people in a room, our robot may be able to harness its expectations about future events to guide its visual processing in top-down fashion.

ACKNOWLEDGMENT

Support for this work was provided by a National Science Foundation CAREER award (#0238334) and NSF award #0534610 (Quantitative Measures of Social Response in Autism). Some parts of the architecture used in this work was constructed under NSF grants #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots) and from the DARPA CALO/SRI project. This research was supported in part by a software grant from QNX Software Systems Ltd.

REFERENCES

[1] G. Sun and B. Scassellati, "Reaching through learned forward model," in *IEEE-RAS/RSJ International Conference on Humanoid Robots*. Santa Monica, CA: IEEE, 2004.

[2] K. Gold and B. Scassellati, "Grounded pronoun learning and pronoun reversal," in *5th International Conference on Development and Learning (ICDL-06)*, 2006.

[3] ——, "Learning about the self and others through contingency," in *AAAI Spring Symposium on Developmental Robotics*. Palo Alto, CA: AAAI, 2005.

[4] M. W. Doniec, G. Sun, and B. Scassellati, "Active learning of joint attention," in *IEEE-RAS International Conference on Humanoid Robotics (Humanoids 2006)*, 2006.

[5] F. Heider and M. Simmel, "An experimental study of apparent behavior," *American Journal of Psychology*, vol. 57, pp. 243–259, 1944.

[6] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in Cognitive Sciences*, vol. 4, no. 8, pp. 299–309, 2000.

[7] H. Choi and B. J. Scholl, "Perceiving causality after the fact: Postdiction in the temporal dynamics of causal perception," *Perception*, vol. 35, pp. 385–399, 2006.

[8] A. M. Leslie, F. Xu, P. D. Tremoulet, and B. J. Scholl, "Indexing and the object concept: developing 'what' and 'where' systems," *Trends in Cognitive Sciences*, vol. 2, no. 1, pp. 10–18, 1998.

[9] B. J. Scholl, "Can infants' object concepts be trained?" *Trends in Cognitive Sciences*, vol. 8, no. 2, pp. 49–51, 2004.

[10] S. R. Mitroff and B. J. Scholl, "Forming and updating object representations without awareness: evidence from motion-induced blindness," *Vision Research*, vol. 45, pp. 961–967, 2004.

[11] P. Wolff, "Representing causation," *Journal of Experimental Psychology*, vol. 136, pp. 82–111, 2007.

[12] L. Talmy, "Force dynamics in language and cognition," *Cognitive Science*, vol. 12, pp. 49–100, 1988.

[13] G. Gigerenzer and P. M. Todd, *Simple Heuristics that Make Us Smart*. Oxford University Press, 1999, ch. 12.

[14] H. C. Barrett, P. M. Todd, G. F. Miller, and P. W. Blythe, "Accurate judgments of intention from motion cues alone: A cross-cultural study," *Evolution and Human Behavior*, vol. 26, pp. 313–331, 2005.

[15] N. B. Priyantha, "The cricket indoor location system," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.