*Brian Scassellati, Christopher Crick, Kevin Gold, Elizabeth Kim, Frederick Shic, and Ganghua Sun*
*Yale University, USA*

# Social Development

© IMAGESTATE

## I. Introduction

Most robots are designed to operate in environments that are either highly constrained (as is the case in an assembly line) or extremely hazardous (such as the surface of Mars). Machine learning has been an effective tool in both of these environments by augmenting the flexibility and reliability of robotic systems, but this is often a very difficult problem because the complexity of learning in the real world introduces very high dimensional state spaces and applies severe penalties for mistakes.

Human children are raised in environments that are just as complex (or even more so) than those typically studied in robot learning scenarios. However, the presence of parents and other caregivers radically changes the type of learning that is possible. Consciously and unconsciously, adults tailor their actions and the environment to the child. They draw attention to important aspects of a task, help in identifying the cause of errors, and generally tailor the task to the child's capabilities.

Our research group builds robots that learn in the same type of supportive environment that human children have and develop skills incrementally through their interactions. Our robots interact socially with human adults using the same natural conventions that a human child would use. Our work sits at the intersection of the fields of social robotics [1], [2] and autonomous mental development [3]. Together, these two fields offer the vision of a machine that can learn incrementally, directly from humans, in the same ways that humans learn from each other. In this article, we will introduce some of the challenges, goals, and applications of this research.

## II. Social Cue Perception

Recognition of certain social cues has been studied extensively in machine vision and pattern recognition. For example, finding faces, detecting people, and tracking body pose are all active research problems for the machine vision community. In this section, we highlight the perception of two social cues that are less well studied but potent components of social learning: perception of vocal prosody and perception of intention from visual motion.

### *A. Perceiving Vocal Prosody*

Vocal prosody can roughly be defined as tone of voice. It is conveyed by intonation, rhythm, and lexical stress in speech.

1556-603X/06/$20.00©2006IEEE

Prosody communicates both paralinguistic information about the affective state of the speaker as well as linguistic information about the intended meaning of spoken utterances. Prosody contributes to learning by providing a natural vehicle for social feedback. Even preverbal infants stiffen in response to prohibitive utterances [4] and tend to smile in response to encouraging tones [5].

**Scaffolding is the the process of using simple capabilities that the agent already has mastered to enable more rapid and/or efficient learning of a complex skill. Social development acts as a powerful tool for scaffolding.**

In signal processing terms, the perception of prosody is a complicated problem, involving measurements of pitch, segmentation of temporal phenomena such as syllable, word and utterance boundaries, and the measurement of stress or energy features. Most work in this area (including some from our own group) focuses on building classifiers that discriminate between prosodic classes such as approval, prohibition, and soothing tones based on extracted acoustic features [6]–[8].

Because generation of prosody is natural and automatic, and because the recognition of prosody is present from infancy, prosody is an attractive method for obtaining feedback for a developmental system. Our group is currently developing machine learning applications which rely on prosodic communication of affect as feedback. Prosody is also influential in learning tasks because it can aid task segmentation. When teaching a task to preverbal infants, adults modulate their prosody to help segment long, complicated instructions into shorter, simpler tasks [9]. No computational systems to date have made use of prosody for segmentation.

### *B. Perceiving Intention from Motion*

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 [10]–[12]. 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.

Within our group, we are working to build systems which automatically analyzes the spatiotemporal relationships of real-world human activity to infer goals and intentions. 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. To date, rather than attempting to solve the complex machine vision problem of segmentation, we have been using a distributed sensing network consisting of radio- and acoustic-enabled nodes. 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 attached to each agent and object involved in a particular dramatic scenario, and eight listener nodes in fixed positions throughout the room, we can determine a person or object's location within a centimeter or two through triangulation.

We then model the movement of a particular agent as a potential field composed of attractive and repulsive forces from each of the other agents and objects in the scenario. From a particular motion trajectory, we generate a set of hypotheses about possible attractive/repulsive relationships and then use least-squares fitting to find the best match. The most interesting events, nearly always noticed and properly interpreted by human observers, occur when an agent's motivations for movement change, and our approach can detect these events. Using simple cues (such as large changes in direction of motion), we segment an agent's trajectory into several sections and compute separate vector field estimates for each segment.

We have validated this methodology by comparing our computational models to human performance. Each spatiotemporal recording is translated into an animation (similar to those of [10]) in order to remove contextual effects. From both free response questions and from matching possible descriptions to a particular animation, most subjects identified the same relationships between the pictured entities and noticed the same dramatic moments where those relationships change. This technique allows us to continue to process only low-level sensor information (visual movement trajectories) in order to obtain simplistic cognitive interpretations of intent and goal.

## III. Using Social Cues to Scaffold

Scaffolding is the process of using simple capabilities that the agent already has mastered to enable more rapid and/or efficient learning of a complex skill. Social development acts as a powerful tool for scaffolding, allowing the child to make use of not only its own competencies, but also to rely on the knowledge and expertise of an instructor. The higher-level skills that are enabled by basic social cue recognition are sometimes social skills, but often are skills that are primarily sensorimotor, cognitive, or communicative. The primitive social skills provide the feedback to the instructor on how well the child understands the current task, provides a support mechanism by which the adult can manipulate the responses of the child, and allows the child to receive feedback directly from the instructor. To illustrate these points,

we provide two examples of low-level social cues providing scaffolding support for higher-level capabilities.

### *A. Learning Joint Attention via Social Scaffolding*

Joint attention refers to the ability to find and look at the same object another person is looking at. Joint attention is not merely a coincidence of two lines of gaze; it is a critical skill for word learning [13] and is hypothesized to be one of the core deficits in pervasive social disorders such as autism [14]. Infants learn to attend to objects of mutual interest over a period of several months, but mastery is not obtained until one and a half years of age.

**Our research group builds robots that learn in the same type of supportive environment that human children have and develop skills incrementally through their interactions.**

It has been suggested that infants can learn joint attention skills simply by guessing [15], [16], that is, when infants think the mother is looking at something interesting, they note her head pose at that particular moment and search the space around her for interesting objects. The position of the first interesting object they find is then associated with the registered head pose. Although this idea has some merits, we believe that joint attention can be learned actively and more efficient way.

We recently used an active learning paradigm to improve the accuracy and learning speed for joint attention behaviors on a humanoid robot [17]. In each experimental trial, six objects (stuffed animals) were placed at predetermined positions between our humanoid robot Nico and an experiment subject (see Figure 1). For each experimental trial, Nico first looked down toward the objects and then looked back toward the subject while pointing to one of the objects with its arm. The response of the subject is recorded by Nico's computer vision system for a few seconds and then Nico again looks down toward the objects.

The video sequences collected from the experiment are processed by a tracker which outputs the head pose of the person in the frame [18]. By tracking the change in head pose, we can estimate when the human has attended to a location and can associate activities of the robot (such as the pointing gesture) with particular head poses of the human. A Radial Basis Function Network (RBFN) trained on one experiment subject's data (100 associations in the form of head pose and object position pairs) can predict with 90% accuracy the object of interest among six possible distracters. If the same RBFN is used to predict the object of interest of other experiment subjects, it is accurate 73.74% of the time (see Figure 2). Perhaps most importantly, this simple social interaction allows our system to obtain superior recognition accuracy with two orders of magnitude fewer training examples than either [15] or [16].

### *B. Social Skills Guide Language Development*

Social cue recognition can also be a fundamental part of the development of language, an observation that runs counter to most purely statistical methods for speech recognition and understanding. Though most speech recognition systems today require a fixed vocabulary drawn from a large corpus, it may eventually be useful to have robots that can learn new words over the course of their operation, and connect them to their sensed environments. By considering how human children learn language, we can potentially avoid costly assumptions about how such a language-learning system should work, and ultimately engineer more flexible systems.

A natural first approach for teaching a robot the meanings of words would be to program the robot to statistically find images and sound sequences that are commonly experienced together [19]. However, this approach fails for the deictic pronouns "I" and "you," two pronouns that children learn quite early. Point to the robot's mirror image and say "This is you," and the robot will have no way of knowing that it should not refer to itself as "you." On the other hand, if the robot associates "I" with only its own image, it will fail to understand humans using the word. How, then, do human children learn these words?

One tool that human children have at their disposal that has been previously under-utilized in robotic implementations is that they appear to learn some words by observing other people interact, rather than through one-on-one interaction with a teacher [20]. By observing the shifting reference of "I" and "you" in a conversation, a child can infer that these words do not refer to particular individuals, but anyone assuming the roles of speaker and listener. This

**FIGURE 1** A humanoid robot (named Nico) uses a learned pointing behavior to acquire joint attention skills. When the robot points to one of the objects lying in front of it, the caregiver will very likely look at what the robot is pointing to. The robot learns to associate spatial positions with head postures and can then extrapolate from a head posture to a location in the world.

principle has proven useful to us in implementing a robotic system that can learn the correct usage of "I" and "you" from observation alone [21].

Though a statistical approach that associates everything in the robot's environment with everything the robot hears is too poorly targeted to be useful, statistical approaches can be highly efficient if the word's referent in a particular instance is already understood, and only the meaning or property being referred to must be inferred. In our word learning system, chi-square tests are used to rank definitions by statistical significance, so that words are more likely to be associated with the most uncommon properties of their referents. Our analysis has shown that using chi-square tests in this way produces values that grow linearly with the number of times a word has been used, and inversely with the frequency with which the property has been observed [22].

As one final grounding in primitive social abilities, we have unified the pronoun-learning system with a system that discriminates self from other using a temporal correlation method [?]. The ability to identify the self in a mirror reflection and the ability to use the word I effectively are commonly seen as major milestones in a human infants development of a concept of self. To discriminate between self and other, the robot learned the timing of the visual feedback that results from its own arm's movement. By noting that there is a characteristic delay between the robot

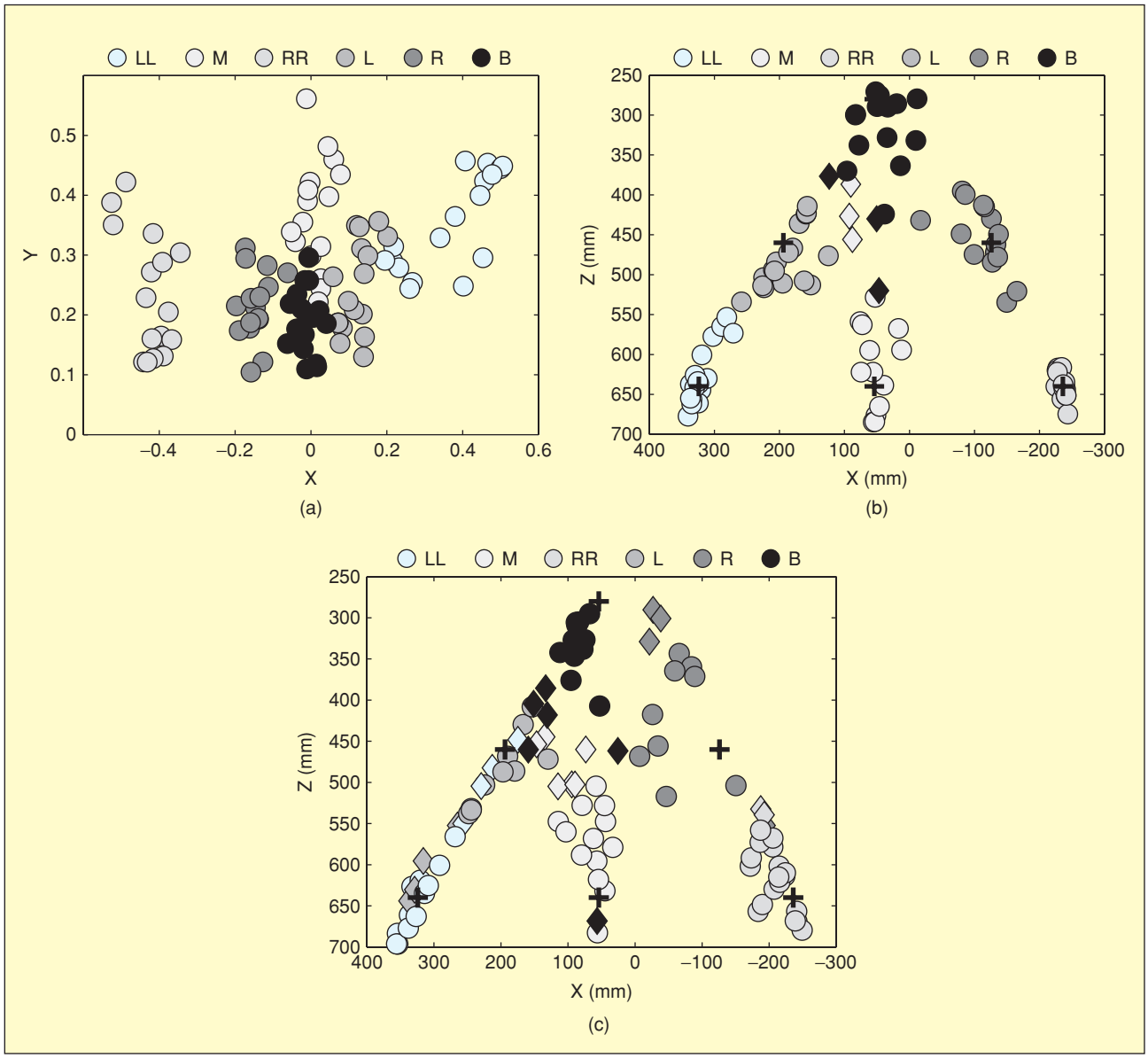


**FIGURE 2** (a) Head pose data of one of the authors (100 samples). Each marker represents a head pose vector projected on the X and Y axis of a horizontal plane extending from the base of the robot. (b) Projected object position data. The RBFN used for this projection is trained with data gathered on the same subject. Out of one hundred samples, only six are misclassified. (c) Projected object position data of the other experiment subjects. The RBFN trained on the main test subject is tested on the hundred data samples gathered on five other test subjects. Although the number of misclassifications is larger, the result is still impressive considering the variations among the five different test subjects.

issuing a motor command to the arm and the act of perceiving the visual movement of the arm, the robot is able to correctly discriminate its own body from others, including when it is seeing not itself directly but rather a reflection of itself in a mirror (see Figure 3). In both cases, the robot recognizes itself as being at a particular location in visual space and can then refer to itself using the correct pronoun and can use that word correctly in context when asked select questions. For example, when trained with a small grammar that involves catching and throwing, the robot can correctly state "**I** caught the ball" when it does so.

**In addition to providing increased performance for computational learning tasks, the construction of perceptual and cognitive systems for allowing robots to interact socially with humans offers a unique tool for characterizing human social development.**

## IV. Characterizing Human Social Development

In addition to providing increased performance for computational learning tasks, the construction of perceptual and cognitive systems for allowing robots to interact socially with humans offers a unique tool for characterizing human social development. Perhaps most importantly, these robots are unique tools in the monitoring and diagnosis of social disorders such as autism.

Autism is a pervasive developmental disorder that is characterized by social and communicative impairments. The social disability in autism is a profound one affecting a persons capacity for understanding other people and their feelings, and for establishing reciprocal relationships. While it is clear that autism is a brain-based disorder with a strong genetic basis, the cause of autism is unknown. Furthermore, autism remains a behaviorally specified disorder [23]; there is no blood test, no genetic screening, and no functional imaging test that can diagnose autism. Diagnosis relies on the clinicians intuitive feel for the child's social skills including eye-to-eye gaze, facial expression, body postures, and gestures. These observational judgments are then quantified according to standardized protocols that are both imprecise and subjective (e.g. [24], [25]). The broad disagreement of clinicians on individual diagnoses creates difficulties both for selecting appropriate treatment for individuals and for reporting the results of population-based studies [26], [27].

Many of the diagnostic problems associated with autism would be alleviated by the introduction of quantitative, objective measurements of social response. We believe that this can be accomplished through two methods: through passive observation of the child at play or in interactions with caregivers and clinicians, and through structured interactions with robots that are able to create standardized social presses designed to elicit particular social responses. While the information gathered from both passive and interactive systems will not replace the expert judgment of a trained clinician, providing high-reliability quantitative measurements will provide a unique window into the way in which children with autism attempt to process naturalistic social situations. These metrics provide both an opportunity to compare populations of individuals in a standardized manner and the possibility of tracking the progress of a single individual across time. Because some of the social cues that we measure (gaze direction in particular) are recorded in greater detail and at an earlier age than can occur in typical clinical evaluations, one possible outcome of this work is a performance-based screening technique capable of detecting vulnerability for autism in infants and toddlers.

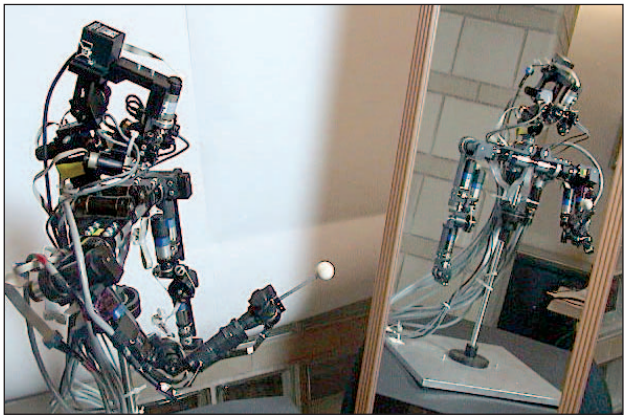

**FIGURE 3** Using a learned temporal filter that characterizes the delay between issuing a motor command and the perceived movement that results from that motor command, the robot Nico has learned to discriminate self from other, including when seeing itself in a mirror. This basic social skill acts as one piece of scaffolding for more complex learning, in this case, learning to use the pronouns "**I**" and "you" correctly.

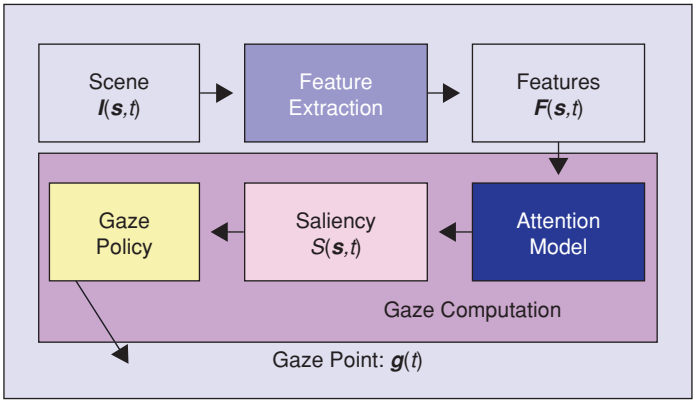


**FIGURE 4** A generic framework for computational models of visual attention. Features *F* are extracted from the input scene *I*, which are then used to determine a saliency map *S* that quantifies the interest level for each spatiotemporal point. This saliency map can then be used to generate a sequence of gaze positions using a particular selection policy.

Since the eye can only focus upon one point in the visual scene, an individual must make a decision as to what to attend to at any given point in time. This process of allocating attention can help us gain insight into the internal cognitive processes of an individual: by watching the eyes, we can extract what is important to that individual at that moment. Furthermore, if we know both what is occurring in the scene as well as where an observer is looking, we can correlate these two systems with one another and extract from their relationship the intrinsic personal saliency of the visual scene to the individual by computational modeling. By comparing models of saliency of one individual to models of saliency of others, we can derive from this data population specific effects.

In our work, we have taken the gaze patterns of individuals with autism and matched controls as these individuals watch scenes from the 1966 black-and-white movie "Who's Afraid of Virginia Woolf?" and analyzed these gaze patterns computationally. By framing visual attention in a simplified computational model (Figure ??), we are able to test the various interactions between the scene and the individual. The scene is first preprocessed using traditional methods for temporal and spatial re-sampling and feature extraction. These features are then operated over by an attention model in order to generate a saliency map. This saliency map is a computational representation that assigns a value of the attractiveness of every spatiotemporal point to an observer. Once we have a saliency map, we can gauge how likely it is that an individual will attend to a specific location in the scene. A gaze policy maps the saliency map to a specific gaze location.

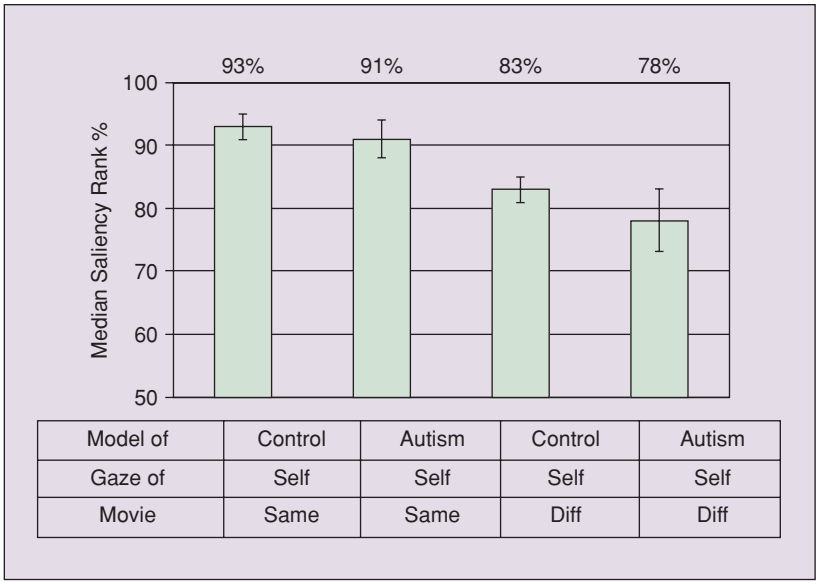


**FIGURE 5** Self-tuning comparisons across movies. Results are aggregated (N = 10 in each condition) for models trained on one individual (control or autism) and tested on the gaze patterns of that same individual (watching either the same movie or a different movie). When a model is trained on one movie and applied to another movie, we get a drop in performance. However, in all cases, human models describe the gaze of other humans much better than random as determined theoretically (50%) and empirically (5213%, N = 600). Error bars span two standard deviations.

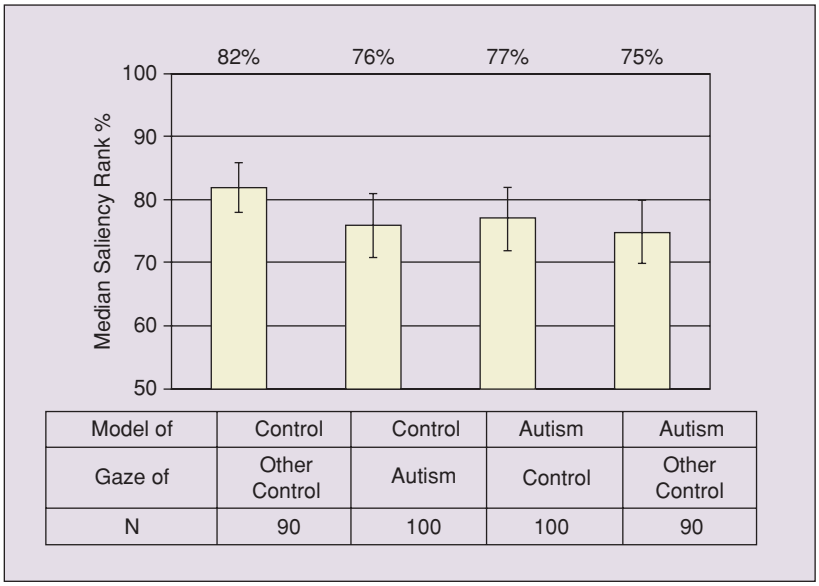


**FIGURE 6** Cross-tuning comparisons within the same movie clip. Models for the gaze of controls describe the gaze of other controls better than the any cross-population comparison that involves autism, including autism models applied to the gaze of other individuals with autism.

In our analysis of individuals with autism and typical controls, we are primarily interested in group effects. For this reason we obtain the saliency maps associated with each subject in our study, and we employ these models in the evaluation of other subjects. In other words, we fit the parameters of the visual attention model to match the gaze patterns of each individual subject. This results in a series of saliency maps for each individual. If we look at the locations that the individual actually attended to, those associated saliency values should be high (they will not, in general, be the highest in the scene, since there are many stochastic steps involved in the actual human gaze decision process, and because modeling is in some way related to estimation). If we look at the points where some other individual attends to in the same scene, and we look at the saliency values associated with the original individual's model, we will find saliency values that are lower. How far apart the values are gives us an indication of how different the scanning strategies of these two individuals are.

In [28], for a given spatiotemporal point $(s, t)$ we used the linearization of the

$11 \times 11$ image patch centered spatially at $s$ and at times $(t - 100\,ms)$ and $(t - 300\,ms)$ as our features $F(s, t)$. The attention model is formed by taking the Fisher's linear discriminant of the locations attended-to by an individual as compared to locations not-attended-to (as obtained by sampling 15 points well separated spatially from the attended-to location). That is, for each individual we determine $w$, the optimal 1D projection for discriminating between attended-to and non-attended-to locations. We then obtain our saliency maps by projecting the features of the scene, i.e., $S(s, t) = wF(s, t)$. By grouping the statistics associated with each individual $i$, $S_i(gi(t), t)$, we can obtain insight as to the underlying distances between individuals with autism and typical controls (Figures 5 and 6).

The application of our framework leads to several results. First, all applications of a human's model to a human's gaze trajectory lead to performance much better than those obtained by random chance (evaluated by synthetic gaze trajectories; $p < 0.01$). This suggests that both individuals with autism and control individuals rely on some common scanning approach, implying the existence some core human strategy. Furthermore, this result suggests that it is unlikely that a methodological bias exists in either the learning technique or the feature representation. Second, the extremely high matched-application (control on self and autism on self groupings) within-movie scores suggest that each subject relies upon some specific individual strategy. This specific individual strategy does not seem to transfer across scenes, as demonstrated by matched comparison score drops as we move from within-movie comparisons to across-movie comparisons, suggesting that top-down or contextual influences on gaze strategy are significant. Third, control individuals, who are taken to be socially more typical than individuals with autism, exhibit much greater coherence ($p < 0.01$) in terms of attraction to underlying features than cross-application cases that involve individuals with autism. This suggests that the strategies of controls transfer well to other controls, but that the strategies of individuals with autism do not transfer to the same degree to either normal individuals or even other individuals with autism.

## V. Conclusion

Research at the intersection of social robotics and autonomous mental development attempts to understand how robots might develop social learning capabilities. We have shown examples of systems that generate and perceive social cues (such as vocal prosody) and that attempt to build cognitive skills from basic perceptual systems. The role of social development as a facilitator for scaffolding can be seen in the enhanced sensorimotor learning performance that is seen with reaching and joint attention behavior and in the methods for enhancing language learning seen in the example of learning "I" and "you." Finally, these same systems that we develop for reasons of computational expediency turn out to be effective tools in the quantitative analysis of social behavior and as diagnostic instruments in the treatment of autism.

## VI. Acknowledgments

Support for this work was provided by a National Science Foundation CAREER award (#0238334), award #0534610 (Quantitative Measures of social response in autism), and awards #0205542 (ITR: A Framework for Rapid Development of Reliable Robotics Software) and #0209122 (ITR: Dance, a Programming Language for the Control of Humanoid Robots). This research was supported in part by a grant of computer software from QNX Software Systems Ltd and a clinical initiatives grant by the Doris Duke Foundation.

## References

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," 2002, [Online]. Available: citeseer.ist.psu.edu/fong03survey.html

[2] C. Breazeal and B. Scassellati, "Robots that imitate humans," *Trends in Cognitive Sciences*, vol. 6, pp. 481–487, 2002.

[3] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," *Science*, vol. 291, no. 5504, pp. 599–600, Jan. 2000.

[4] M.M. Lewis, *Infant speech: A study of the beginnings of language*, London: Routledge & Kegan Paul, 1936.

[5] A. Fernald, "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?" *Child Development*, vol. 60, pp. 1497–1510, 1989.

[6] A. Robinson-Mosher and B. Scassellati, "Prosody recognition in male infant- directed speech," in *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pp. 2209–2214, 2004.

[7] C. Breazeal and L. Aryananda, "Recognition of affective communicative intent in robot-directed speech," *Autonomous Robots*, vol. 12, pp. 83–104, 2002.

[8] M. Slaney and G. McRoberts, "Baby ears: A recognition system for affective vocalizations," in *Proc. of 1998 Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[9] F.J. Wrede, B. and K. Rohlfing, "How can prosody help to learn actions?," in *Fourth IEEE International Conference on Development and Learning (ICDL)*, 2005.

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

[11] B. Reeves and C. Nass, *The media equation: how people treat computers, television, and new media like real people and places*, New York, NY, USA: Cambridge University Press, 1998.

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

[13] D.A. Baldwin, "Understanding the link between joint attention and language," in *Joint Attention: Its origins and role in development*, C. Moore and P.J. Dunham, Eds., Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 131–158, 1995.

[14] S. Baron-Cohen, *Mindblindness: Ann essay on autism and theory of mind*, Cambridge, MA: MIT Press, 1995.

[15] J. Triesch, C. Teuscher, G. Deak, and E. Carlson, "Following: why (not) learn it?," *Developmental Science*, vol. 9, no. 2, pp. 125–147, 2006.

[16] Y. Nagai, "Understanding the development of joint attention from a viewpoint of cognitive developmental robotics," Ph.D. dissertation, Osaka University, 2004.

[17] M. Doniec, G. Sun, and B. Scassellati, "A demonstration of the efficiency of developmental learning," in *Proceedings of the 2006 International Joint Conference on Neural Networks*, 2006.

[18] L. Morency, A. Rahimi, and T. Darell, "Adaptive view-based appearance model," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

[19] D.K. Roy and A.P. Pentland, "Learning words from sights and sounds: a computational model," *Cognitive Science*, vol. 26, pp. 113–146, 2002.

[20] Y. Oshima-Takane, Y. Takane, and T. Shultz, "The learning of first and second person pronouns in English: network models and analysis," *Journal of Child Language*, vol. 26, pp. 545–575, 1999.

[21] K. Gold and B. Scassellati, "Using context and sensory data to learn first and second person pronouns," in *Human-Robot Interaction 2006*, Salt Lake City, Utah, 2006.

[22] ——, "Grounded pronoun learning and pronoun reversal," in *International Conference on Development and Learning*, Bloomington, IN, 2006, under review.

[23] F. Volkmar, C. Lord, A. Bailey, R. Schultz, and A. Klin, "Autism and pervasive developmental disorders," *Journal of Child Psychology and Psychiatry*, vol. 45, no. 1, pp. 1–36, 2004.

[24] E. Mullen, *Mullen Scales of Early Learning*, ags edition ed., Circle Pines, MN, 1995.

[25] S. Sparrow, D. Balla, and D. Cicchetti, *Vineland Adaptive Behavior Scales*, ags edition ed., Circle Pines, MN, 1984.

[26] A. Klin, J. Lang, D. Cicchetti, and F. Volkmar, "Interrater reliability of clinical diagnosis and dsm-iv criteria for autistic disorder: Results of the dsm-iv autism field trial," *Journal of Autism and Developmental Disorders*, vol. 30, no. 2, pp. 163–167, 2000.

[27] F. Volkmar, K. Chawarska, and A. Klin, "Autism in infancy and early childhood," *Annual Review of Psychology*, vol. 56, pp. 315–36, 2005.

[28] F. Shic and B. Scassellati, "A behavioral analysis of robotic models of visual attention," *International Journal of Computer Vision*, 2006, accepted, to appear.