

# **Bidirectional gaze guiding and indexing in human-robot interaction through a situated architecture**

by

Nicholas Brian DePalma

M.Sc. Computer Science, Georgia Institute of Technology (2010)

B.Sc. Computer Science, Georgia Institute of Technology (2005)

Submitted to the Program in Media Arts and Sciences  
School of Architecture and Planning  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the


MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology, 2017. All rights reserved.

  
**Signature redacted**

Author .....

  
Program in Media Arts and Sciences  
School of Architecture and Planning

**Signature redacted** June 1, 2017

Certified by .....

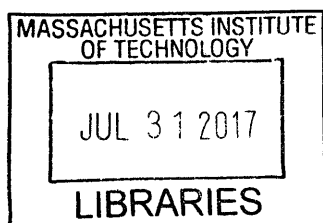
Dr. Cynthia Breazeal  
Associate Professor of Media Arts and Sciences  
Thesis Supervisor

  
**Signature redacted**

Accepted by .....

Dr. Pattie Maes  
Academic Head

Program in Media Arts and Sciences



ARCHIVES



# **Bidirectional gaze guiding and indexing in human-robot interaction through a situated architecture**

by

Nicholas Brian DePalma

Submitted to the Program in Media Arts and Sciences  
School of Architecture and Planning  
on June 1, 2017, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy

## **Abstract**

In this body of work, I present a situated and interactive agent perception system that can index into its world and, through a bidirectional exchange of referential gesture, direct its internal indexing system toward both well-known objects as well as simple visuo-spatial indexing in the world. The architecture presented incorporates a novel method for synthetic human-robot joint attention, an internal and automatic crowdsourcing system that provides opportunistic and lifelong robotic socio-visual learning, *supports the bidirectional process of following referential behavior*, and *generates referential behavior useful for directing the gaze of human peers*. This document critically probes questions in human-robot interaction around our understanding of gaze manipulation and memory imprinting on human partners in similar architectures and makes recommendations that may improve human-robot peer-to-peer learning.

Thesis Supervisor: Dr. Cynthia Breazeal

Title: Associate Professor of Media Arts and Sciences





**Bidirectional gaze guiding and indexing in human-robot interaction  
through a situated architecture**

by

Nicholas Brian DePalma

M.Sc. Computer Science, Georgia Institute of Technology (2010)

B.Sc. Computer Science, Georgia Institute of Technology (2005)

Submitted to the Program in Media Arts and Sciences  
School of Architecture and Planning  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology, 2017. All rights reserved.

**Signature redacted**

Certified by ...

.....  
Dr. Brian Scassellati  
Professor of Computer Science, Mechanical Engineering,  
and Material Science  
Yale University  
Thesis Reader

**Signature redacted**

Certified by .....

.....  
Dr. Julie Shah  
Associate Professor of Aeronautics and Astronautics  
Massachusetts Institute of Technology  
Thesis Reader



## Acknowledgments

When I first decided that robotics was the next big thing in 2006, I couldn't have imagined how the first eight years of this journey would transpire. While I have found myself intellectually interested in the problems of interaction and more generally sociability in artificial intelligence, I have learned much to inform my world view from other pillars of the artificial intelligence and cognitive science spectrum. There have been so many individuals and organizations that have helped me understand the research and development community. I want to thank just a few of these individuals.

Foremost to my experience, Dr. Andrea Thomaz took a risk when she welcomed a curious computer vision engineer into her research group and I will always be grateful for that. When I first began my intellectual journey at Georgia Tech, I immediately targeted her group and tried to work my way into her good graces. Her view of robotics and sociability helped shape the foundation of all I believed possible in robotics. To this day, I still have ideas I believe are novel and then find in a literature review that her lab has already had this idea or has explored it in an interesting way. Along with Dr. Thomaz, Maya Cakmak, Crystal Chao, Chien-Ming Huang, and Michael Gielniak provided a friendly environment for this aspiring researcher. The professors and citizens of the Georgia Tech Robotics and Intelligent Machines center provided an enriching environment in which to grow. Thanks to Dr. Charles Isbell for his humor, Dr. Henrik Christensen for his impressive breadth and depth of robotics knowledge and of course Dr. Mike Stillman. Mike demonstrated the true conviction and dedication of a roboticist. He helped me through my masters thesis and gave me the set of tools to persevere through a research program. Additionally, he had an amazingly prescient and quick answer for one of my biggest questions that still eats away at me: what is the minimal hardware needed to study robotics? His response that robotics can just be an arm, some wheels, and a sensor array still represents one of my intellectually solid baseline answers to this day for researchers interested in mobile manipulation.

When I transitioned from Andrea's lab, I had some idea of the challenges I would face in Dr. Cynthia Breazeal's lab. I have never encountered someone so full of ideas and breadth of understanding of so many topics in robotics. Cynthia's ideas are about 30+ years

in the future and you truly can't help but be a believer once you see the world through her eyes. I want to wholeheartedly thank Cynthia for my time in her lab and her support of my intellectual pursuit. I cannot emphasize how much I matured in such a short time due to her timely responses to a number of my intellectual dead-ends. Additionally, I want to thank her for providing an attentive ear for blue sky ideas and discussions. I know some of my own were half baked or led to dead ends, but she helped them along and I'll always appreciate that.

I will forever be humbled to have been a part of the Personal Robots Group. Jin Joo was my buddy during many late nights of software and hardware modifications. I am thankful to have been able to work along side such a stellar researcher who had a critical eye in dozens of domains. Siggi Adalgeirsson was my officemate for four years and had such an amazing eye for *quality* that I will probably have a little Siggi voice in my head helping me understand how to improve my work for the rest of my life. Adam Whiton, Peter Schmitt, Jesse Gray and Angela Chang were from another generation and time but one I constantly looked to for inspiration. To the younger and more successful generation: Sooyeon Jeong, Jackie Kory, and Sam Spaulding, I'll miss you guys and our brainstorming. The post-docs anchored me many times when required and I will always be thankful for their support. Dr. Sonia Chernova helped me understand the facets that made for an interesting idea, Dr. Brad Knox helped me understand the importance of cultural vocabulary and Dr. Goren Gordon pulled me up when I needed it most, sharing both a passion for artificial neural networks and board games. Dr. Hae Won Park joined the group as a post-doc in my last year and while our time working in proximity was short, she provided me with critical feedback when I needed it. I'm just sorry I won't be around for her entire tenure at PRG. Lastly, I want to acknowledge much of the hard work put in by my editors, artists, advisors and undergraduate researchers who have worked with me throughout my tenure. Big thanks to Philip Graham, Elisabeth Morant, Pamela Jones, Will Jobst, Jason Wiser, Yasmin Chavez, Fardad Faridi and Sabrina Shemet.

The support staff of MIT and the Media Lab provided me with the emotional and intellectual support I would need to survive. I want to thank Linda Peterson who cared for the *holistic me* and Keira Horowitz for making me laugh and smile despite the challenges.

I want to thank Polly for her wisdom and advice. I'll never forget her support through the ups and downs of my tenure as a research assistant. The people of the MIT sailing pavilion for letting me sail their boats to the middle of the Charles River to quietly think and Misha Novitsky for teaching me to have fun in the middle of the storm. Thanks to Dr. Sajit Rao for his patience and structure of thought. Through the years, his advice, packaged long ago in conversation, became more salient. Finally my thesis committee provided incredibly constructive ideas and guidance through many long years. I want to thank Dr. Brian Scasselatti and Dr. Julie Shah most of all for their thoughts on my manuscripts throughout my time as a Ph.D. student.

None of this, of course, would have been possible without the support of my family. My wife and life long partner Katie met me in a middle school cafeteria long ago and was probably the only one who thought it cool that I was reading a C++ book at that point in my life. She has been there for many of my toughest moments in life and has stuck it out with me, I will love her and appreciate her for the rest of my life. My beautiful and intelligent daughter Sasha was born right in the middle of my preliminary exams and has been the smiling, joking tickle monster providing the much needed levity in my life. I have incredible gratitude for the parents and grandparents who have helped support us: Darrin Jones, Pamela Jones, Susanna DePalma and Nicholas DePalma Sr. My parents supported me in so many ways throughout the years but the most important thing they have done for me was to let me make my own decisions in life and to provide me with a cheerleading team when I needed it. Thank you so much for being my biggest ally.

Dedicated to Katie & Sasha.



# Contents

<b>Abstract</b>	<b>3</b>
<b>1 Introduction</b>	<b>27</b>
1.1 Motivation . . . . .	30
1.2 Objectives . . . . .	31
1.2.1 Stance and statement . . . . .	31
1.3 Contributions . . . . .	38
<b>2 Robotic architecture overview</b>	<b>45</b>
2.1 Features of SHARE . . . . .	46
2.2 Loop 1: Joint attention for situated robotic interaction . . . . .	47
2.3 Loop 2: On-the-fly crowdsourcing infrastructure . . . . .	49
2.4 Architecture layout and pipeline . . . . .	51
2.5 A core mechanism to improve contribution quality . . . . .	53
2.5.1 Problem domain: on-the-fly visual labeling . . . . .	54
2.5.2 Data collection . . . . .	56
2.5.3 Results: Time-to-response measurements . . . . .	57
2.5.4 Results: Analysis of quality and speed . . . . .	58
2.6 Extended cloud architecture . . . . .	59
2.7 Designing an interactive perception-action system for shared visual atten- tion: next steps . . . . .	62
<b>3 Human-robot shared attention : metric and analysis</b>	<b>69</b>

3.1	Relevant work . . . . .	71
3.2	Shared attention : metric and hypothesis . . . . .	73
3.2.1	A new metric for sharing attention . . . . .	73
3.2.2	Human dyad pilot study . . . . .	74
3.3	Analysis of behavior . . . . .	76
3.4	Discussion . . . . .	77
<b>4</b>	<b>Design of a directable artificial visual search architecture</b>	<b>81</b>
4.0.1	Simulating a peer's reference to improve exophoric resolution . . .	82
4.0.2	Conditions and hypotheses . . . . .	83
4.1	Related work . . . . .	84
4.1.1	Joint attention and attention in robotics . . . . .	84
4.1.2	Shared attention through a common representation: deictics . . . .	86
4.2	Problem: the socially driven visual search task . . . . .	87
4.2.1	Task requirements . . . . .	88
4.3	Computational model . . . . .	90
4.3.1	Competition algorithm . . . . .	91
4.3.2	Interpreting object-directed, referential action . . . . .	92
4.3.2.1	Proposer 1: Reference-to-object effects . . . . .	92
4.3.2.2	Proposer 2: Reference-to-foreground effects . . . . .	94
4.3.3	Generating referential action . . . . .	94
4.3.4	Fan-in unification . . . . .	95
4.4	Data collection . . . . .	96
4.5	Results . . . . .	98
4.5.1	Analysis with competition . . . . .	99
4.6	Discussion . . . . .	101
<b>5</b>	<b>Reference following and object recognition with deictic wells</b>	<b>107</b>
5.0.1	Theory and hypothesis . . . . .	109
5.1	Related Work . . . . .	110
5.1.1	Deep recurrent models of attention . . . . .	110



5.1.2	Joint attention in agent learning . . . . .	111
5.2	Model design . . . . .	111
5.2.1	The baseline RAM model . . . . .	113
5.2.1.1	Reinforcement learning . . . . .	113
5.2.1.2	Reward signal . . . . .	114
5.2.2	The extraction window and deictic observation . . . . .	114
5.2.3	Training the network . . . . .	116
5.3	Data collection and experiment . . . . .	119
5.4	Results and discussion . . . . .	120
5.4.1	Elapsed time in epochs . . . . .	120
5.4.2	Deictics speed up convergence . . . . .	120
5.4.3	Learned policy performance . . . . .	122
5.5	Summary . . . . .	123

## 6 Sustainability of longitudinal human-robot joint attention

	<b>in dynamic environments: adaptation and imprinting</b>	<b>129</b>
6.0.1	Theory and hypothesis . . . . .	130
6.1	Related work . . . . .	132
6.1.1	Deictics in robotics . . . . .	133
6.1.2	Visual attention in robotics and vision . . . . .	133
6.2	Computational model . . . . .	134
6.3	Study setup, details and recruitment . . . . .	138
6.3.1	Experimental conditions . . . . .	140
6.3.2	Novel quantitative metrics . . . . .	141
6.3.2.1	Measuring the effect of social action on a participant's gaze trajectory: a novel metric . . . . .	142
6.3.2.2	Measuring a participant's saliency: a novel metric . . . . .	142
6.4	Results . . . . .	142
6.4.1	The effect of social action . . . . .	143
6.4.2	Effects of social action: a saliency centered analysis . . . . .	146

6.4.3	Memory and performance . . . . .	147
6.5	Discussion . . . . .	150
6.5.1	Unexpected factor: action staging . . . . .	151
<b>7</b>	<b>Reflections and directions</b>	<b>157</b>
7.1	Contributions . . . . .	157
7.2	Publications . . . . .	160
7.3	Recommendations for future work . . . . .	161
7.4	Final remarks . . . . .	165
	<b>Appendix</b>	<b>169</b>
<b>A</b>	<b>Datasets used throughout the dissertation</b>	<b>169</b>
A.1	The tangram dataset (snapshot October 15, 2015) . . . . .	170
<b>B</b>	<b>Recruitment documentation and consent forms</b>	<b>173</b>
B.1	Consent forms . . . . .	174
B.2	Instructions provided to participants . . . . .	181
B.3	Questionnaires . . . . .	183

# List of Figures

1-1	Situating the NIMBUS subsystem within the larger architecture (SHARE).	29
2-1	a) Long term vision of longitudinal embodied visual learning, b) system architecture and design of the NIMBUS architecture: sequence diagram, server subsystem and caching mechanisms. . . . .	51
2-2	Data collection, conditions and user interface for crowdsourcing labels for a situated robotic agent. . . . .	55
2-3	Elapsed times of the two conditions: with and without the novel rollover mechanism. Response times are significantly different ( $p < 0.01$ ) with one-shot labels outperforming rollover. . . . .	57
2-4	Extending the system architecture for use within the rest of the dissertation. a) Extended system architecture as a metaphor for ongoing learning systems. The goal of the extended system is to enable both on-demand and longitudinal iteratively improving algorithms relying on on-line social feedback. Ongoing monitoring of the constantly running system b) in production mode and c) in sandbox mode. Sandbox mode enables testing prior to production deployment. The overall system has been running for 3+ years without stop. . . . .	63

3-1	Observed deictic action and gesture during pilot. Tangrams offer a constrained domain for both selection and recognition. a) Demonstrates an example of a tangram. The observed gesture throughout the experiment was diverse. b) In this excerpt screenshot, iconic gesture (a rarely used referential act) was observed. Another gesture observed throughout the experiment can be found in c) where single extension, sweeping gesture were observed and d) where numerous, precise gestures were observed. . . . .	74
3-2	An analysis of deictic exchange: action class and numeracy. a) Dyad preference for reference strategy. X-axis: Scene (T1-4) Y-axis: +1: 100% general gesture, 0: equal gesture types utilized, -1: 100% precise gestures used. Each line represents a dyad (D1-5). b) Of those dyads that correctly communicated the foreground, this chart presents the fraction of participants who used the basic referencing policy. . . . .	76
4-1	An illustration of synthetic and natural visual search. a) The SHARE system: tangrams in view with foreground being selected and identified as a known “whole object.” The dataset was collected from on-line contributions. Functions available are “find,” “identify” and “classify.” b) A demonstration of the reader’s human visual search adapted from Wolfe (1994). The objective is to find the “T” symbol within the image. Notice how your eye searches the image in no particular order. . . . .	85
4-2	Some observed behaviors during a pilot experiment when attempting to direct the attention of another human participant. Left: a participant used bounded hand gestures to refer to the space between the palms (highlighted in blue). Right: precise pointing meant to highlight one particular region that must be interpreted to be one piece of the tangram figure (highlighted in blue). . . . .	88
4-3	An illustration of capturing the deictic reference to use for selection and hypothesis enumeration. . . . .	90

4-4	Illustrating hypothesis roll-out of known and recognizable wholes from a reference. Horizon of roll-out is given by how proximal or distal the reference is to the location. . . . .	92
4-5	Computing the reference point against a surface from a distal location. . . .	94
4-6	On-line and real world data collection used for collecting recognizable wholes and in person gesturing. Top row: a) Tangram collection web interface. b) Tangram figure taken from the dataset collected on-line. Middle row (both captured from the dataset): c) Object-based foreground goal collected on-line. d) Pixel-based goal foreground collected on-line. Bottom row: e) Pepper's ghost illusion used to situate the tangrams between the participant and the robot. f) Lower portion of the illusion device that uses a Leap Motion mounted on left side to collect gestures toward the illusion. . .	97
4-7	Foreground prediction performance of the visual search mechanism compared against multiple observed action strategies by the participant. The results show that the designed visual attention system performs well while resolving goal foregrounds which the system has already encountered and can recognize (known goal). However, it does not predict the foreground goal well for objects it has never encountered (unknown goal) for both action strategies. Reported significance in this figure are $p < 0.01$ but refer to Table 4.1 for more detail. . . . .	99
4-8	System performance against each test dataset. System-indexical: Proposer pump 1 returns predicted foregrounds $F^P$ based on recognized objects. System-Saliency: Proposer pump 2 returns predicted foregrounds based primarily on the projection of deictic action on the scene itself, $F^P$ . Competition: Resolving how to route each of the sub-pumps in the system to select the correct foreground. In this figure, * refers to significance of $p < 0.05$ while ** refers to significance of $p < 0.01$ . Refer to Table 4.2 for further detail. . . . .	100

5-1	Summary of the RAM model. (a) An example taken from the crowdsourced dataset that resembles and is labeled “cat” (top). The basic extraction network uses the previously computed location position, $l_t$ , and the image data at time, $I_t$ (bottom). (b) The observed state, $f_t$ , is used to train the hidden recurrent network which in turn predicts a reward, then chooses the next best action to take and the next location for moving the artificial fovea. . . .	113
5-2	The proposed <i>dRAM</i> model. (a) Basic extraction that additionally computes the gradient $\delta_w$ and a reward $r_t$ toward an observed, social deictic pointer, $l_t^s$ , at time $t$ . (b) The modified architecture that takes the gradient into account in its hidden state $h_t$ . . . . .	114
5-3	Full dRAM architecture with deictic sampling. (a) Profile of a deictic well (the component that is summed with the previous reward function), (b) the full artificial neural architecture. $l_t^t$ is sampled uniformly from a set of pointing gestures collected from human participants interacting with the robot ( $L_t^S$ ). . . . .	116
5-4	Data Collection. (a) The crowdsourced data collection system. On-line participants can construct tangrams using a mouse and keyboard (left). Highlighted sections (right) are used to select the foreground component and a label can be sourced on-line (e.g., “hull of boat” in this example). This allows us to collect subfigure labels for many different tangram figures as well as full figures. (b) A few sample tangrams collected from the dataset online. . . . .	118
5-5	Test accuracy progress over each epoch plotted with 7 groups. The baseline RAM model converges at a slower rate than the extended model, dRAM. Each dRAM subscript represents the probability of observing a deictic action from the dataset. It is clear that more social interaction has a positive effect on this model. . . . .	121

5-6	Performance of the RAM Model after 100 observation steps. Vertical axis is the model's performance from the 10 test simulations, each row being a particular simulation. (a) The given scene with the fixation point overlaid on top, (b) The extracted patch used by the algorithm. Discussion of these results can be found in Section 5.4.3. . . . .	123
5-7	Performance of the dRAM Model after 100 observation steps. Vertical axis is the 10 learned models, dRAM <sub>1-10</sub> . (a) The given scene with the fixation point overlaid on top. (b) The extracted patch used by the algorithm. Discussion of these results can be found in Section 5.4.3. Note that the observed patch is overwhelmingly sampling from the figure itself after 100 observations in the dRAM model vs. the baseline RAM model (Figure 5-6). 124	
6-1	Synthetic attention architecture component for interaction. Experimental section marked in blue and the attentional components highlighted in green. $z_t$ represents the recognized deictic action from the Leap Motion as a binary event ( $z_t \in \mathbb{Z}$ ), $M_t^T$ represents the trajectory (T) model (M) that returns the location ( $\bar{x}_t = E[x]$ ) as a function of the dataset collected from the NR condition. The [1] function is documented as the NEXTPOSITIONSALIENCY function while the [2] function represents a winner-take-all mechanism (see WTAFOVEA) in Algorithm 6.1. . . . .	135
6-2	Experimental setup with the robot. The robot is an approximately 3.5 foot tall root with fully actuated hands designed to interact with a tilted screen that displays a dynamic narrative. A Leap Motion is mounted just below the display to capture hand gestures. The visual narrative is time synchronized with the robot's deictic action to measure the effect of a particular gesture across the narrative and across the participants. Gaze direction is computed using the method documented in Section 6.2. . . . .	138

6-3	Conditions from left to right: No robot (NR) condition, non-joint attention behavior condition (NJA) and joint attention condition (JA). The arrows represent the effect that the factor has on subject. For this experiment, neither the robot nor the human can affect conspicuity outcomes on the content itself. . . . .	140
6-4	Goal target locations the participant was directed toward at various scenes of the visual narrative (annotated with purple). Notice the gaussian in the JA condition is much closer to the directed location compared to the NJA condition. . . . .	143
6-5	A bar chart of qualitative conspicuity of the social (JA) vs. non social (NJA) conditions as reported by the participants. Participants were able to tell a clear difference between the robot that responded to their deictic action and in return, attempted to direct them to the scene at important moments. Significance reported as Student's <i>t</i> -test. . . . .	144
6-6	Sustainability of Joint Attention: Participants performed reacted in a significantly more timely manner after interaction with the joint interaction robot than with the robot without joint attention. Significance was computed with a Student's <i>t</i> -test. Additionally, it is observed that there is a clear trending reaction time improvement over time for both conditions. Finally, participants were more engaged with or chose to engage with the robot more in the joint attention condition. This resulted in fewer missed probes during the joint attention interaction over time and a more sustained interaction. The number of probes missed in the non-joint attention condition increased over time. Significance cannot be reported due to how the misses were computed which resulted in too few samples in both conditions. . . . .	145



6-7	The participant's mean gaze target (purple) overlaid on top of the saliency map. The purple over the little girl's dress represents the mean gaze target at the moment the robot referenced the auxiliary character on the left. It is clear that saliency of the target region was low at this moment in time compared to the rest of the scene, suggesting that social effects can override the intrinsic conspicuity of the target region. . . . .	146
6-8	Results of human memory recall following study in both the JA and NJA conditions. Unfortunately, no significance was found between the two conditions for any question despite clear evidence that most participants clearly fixated and <i>saw</i> the answer. . . . .	148
6-9	An analysis of the performance of the participant's conditional on how they answered the questions in the post-questionnaire. Purple represents those participants who answered the questions at the end of the experiment correctly while orange represents those who answered incorrectly. Top row represents each condition: a) resulting analysis of response times conditional on answering the question correctly. [NJA condition] b) resulting analysis of response times conditional on answering the question correctly. [JA condition] <i>Note:</i> Top row's N is too small for significance tests, leading to the aggregate analysis in c. c) the aggregate case of NJA+JA conditional on the participant answering correctly. Notice the tighter coupling from those who answered correctly in the post-questionnaire. Significance was computed using a Student's <i>t</i> -test. . . . .	149
A-1	Sample 1 of the tangram dataset captured from participants on-line. Both whole figures are shown here as well as the subfigures that were selected and labeled by the same participants. . . . .	170
A-2	Sample 2 of the tangram dataset captured from participants on-line. Both whole figures are shown here as well as the subfigures that were selected and labeled by the same participants. . . . .	171

A-3	Sample 3 of the tangram dataset captured from participants on-line. Both whole figures are shown here as well as the subfigures that were selected and labeled by the same participants. . . . .	172
B-1	Page 1 of the consent form used throughout the experiments discussed in Chapters 2 and 4. This consent form was covered under COUHES protocol number 1402006168. The protocol was titled “Using Crowd-Sourcing to Understand Effects of Narrative on Perception in Human-Robot Interaction.”	174
B-2	Page 2 of the consent form used throughout the experiments discussed in Chapters 2 and 4. This consent form was covered under COUHES protocol number 1402006168. The protocol was titled “Using Crowd-Sourcing to Understand Effects of Narrative on Perception in Human-Robot Interaction.”	175
B-3	Page 3 of the consent form used throughout the experiments discussed in Chapters 2 and 4. This consent form was covered under COUHES protocol number 1402006168. The protocol was titled “Using Crowd-Sourcing to Understand Effects of Narrative on Perception in Human-Robot Interaction.”	176
B-4	The webpage presented to participants on-line who have already participated in the study. . . . .	177
B-5	Page 1 of the consent form used throughout the experiments discussed in Chapters 4 and 6. This consent form was covered under COUHES protocol number 1403006294. The protocol was titled “Dynamics of Social Attention and Cooperation with Humans”. . . . .	178
B-6	Page 2 of the consent form used throughout the experiments discussed in Chapters 4 and 6. This consent form was covered under COUHES protocol number 1403006294. The protocol was titled “Dynamics of Social Attention and Cooperation with Humans”. . . . .	179
B-7	Page 3 of the consent form used throughout the experiments discussed in Chapters 4 and 6. This consent form was covered under COUHES protocol number 1403006294. The protocol was titled “Dynamics of Social Attention and Cooperation with Humans”. . . . .	180

B-8	The instructions provided to the participant dyad prior to exchanging gesture for the study discussed in Chapter 3. . . . .	181
B-9	The instructions provided to the participants interacting with Maddox for the study discussed in Chapter 4. . . . .	182
B-10	Page 1 of the questionnaire administered to the participants interacting with Maddox following the study discussed in Chapter 6. . . . .	183
B-11	Page 2 of the questionnaire administered to the participants interacting with Maddox following the study discussed in Chapter 6. . . . .	184



# List of Tables

2.1	Quality of response between one-shot labeling and labeling with rollover reported as mean squared error and standard deviation with respect to baseline. Rollover outperforms one-shot labeling in terms of accuracy of response.	59
4.1	Reported significance and error of a single foreground proposer method against known goals and unknown goals. Significance values are reported using a Student's <i>t</i> -test, and the average normalized mean squared error is reported for each dataset on the far right.	99
4.2	Reported significance and error of all subsystems with respect to the dataset. This table is separated into the dataset (by color) and the subsystem (Pump <i>X</i> + Unified). Significance values are reported using a Student's <i>t</i> -test, and the average normalized mean squared error is reported to three decimal places.	100



# Chapter 1

## Introduction

As interactive robots develop, one of the most consistently important behaviors is the ability to follow gaze and gesture in the robot's visual field. This behavior is called *joint attention*, one of the most challenging problems in social and developmental science. Joint attention is traditionally defined as the tendency of humans and some animals to align their focus of attention toward the same stimulus to enable social learning. Joint attention is one of the core cognitive mechanisms that must be understood before robots or other agents can hold meaningful, *transparent* verbal and nonverbal dialogue with a social partner (Seemann, 2011; Kaplan & Hafner, 2006). Transparency is the ability of a machine to reflect on its internal state and report about it to the user of the machine (Dix, 2003). To understand why joint attention is critical in nonverbal and verbal communication, it is instructive to explore the meaning researchers attribute to the word *attention*.

Without a clear understanding of attention and how it works, a robotic system may be misdirected when predicting changes in the attention of a human partner. The capacity that humans have to attribute beliefs to other agents and to simulate the impact of those beliefs on behavior has been called *theory-of-mind*. Theory-of-mind inspires researchers in robotics and cognitive science to better understand and build systems to imitate these capabilities in artificially intelligent systems. *Jointly attending* to the same stimulus is one core way of aligning the *affordances* (Gibson, 1979) (i.e., the actions an agent perceives it can perform upon an object) the robot would jointly infer with a human partner.

One of the largest topics in social robotics is to plausibly represent a theory-of-mind.

Theory-of-mind is our ability to apply beliefs to others and to simulate the thought processes of others. This mechanism allows us to predict subsequent actions and goals of others while making inferences about others' current states-of-mind. This cognitive mechanism has inspired research in self-as-simulator cognitive architectures (J. V. Gray, 2010) as well as research in modifying sequential planning algorithms for use in theory-of-mind contexts (Adalgeirsson, 2014). More specifically, this dissertation takes the theory-of-mind approach as being visually directed and socially coupled.

This dissertation examines differing ideas regarding the concept of joint attention, dissects them and then shows a new path toward a deeper understanding of sharing attention with a social partner. This concept is inspired by the idea that *synthetic computational attention systems* research must be unified with *deictic systems* research at the sensory and motor systems level before situated *common ground* can be addressed. The body of work reported in this dissertation builds an underlying deictic system that can *point* and bind to an underlying visual stimulus while simultaneously attempting to move its internal perception pointer (Ballard, Hayhoe, Pook, & Rao, 1997) closer to a social partner's gaze and instruction. Uniquely, this system operates primarily on visual stimuli captured from a shared environment and provides new, commonly used lexical references from an on-line management subsystem (NIMBUS). The System for Highly Attentive Robotic Experiences or SHARE, as seen in Figure 1-1, offers an environment that unifies image processing and analysis for experimentation, both in human-robot interaction and through browser-based interactions.

One key problem in attention-centered vision systems is the problem of foreground selection. SHARE makes this problem explicit: how can a robot be directed to perceive or learn about the external stimuli using dynamic selection? When a situated agent is directing its perception within its environment, the situated agent can direct the locus of attention to any stimulus, recognizable or not. This type of directed perception has been called *visual attention* in computer vision literature (Rao, 1998; Tang, Srivastava, & Salakhutdinov, 2014; Frintrop, 2006; Bridewell & Bello, 2015; Tsotsos, 2011) and is the only behavioral measure that can be studied related to attention. The direction of the eye, or in other words, the directed perception of the human, is called *gaze* in interaction literature. Neurological



evidence suggests that while gaze does not provide sufficient evidence of which stimuli are being perceived by the agent, it instead hints at features that are not perceived through a mechanism called peripheral gating (Desimone & Duncan, 1995). *Gaze following* emerges in infancy and is shown to be one precursor to sharing attention (Butterworth & Cochran, 1980).

One subtle but important point that distinguishes gaze following from joint attention is joint attention is intentionally driven by interaction with a social partner whereas gaze following is entirely self-directed. In gaze following, no explicit social action is taken to direct a social partner. On the other hand, in joint attention, one agent attempts to direct the attention of another through a *deictic gesture*, or gesture such as pointing, that is intended to guide a social partner. Chen Yu and Linda Smith argue that joint attention unifies visual perception and deictic action, suggesting that these systems may be deeply coupled (Yu & Smith, 2016). This dissertation focuses on the unification of the sensory system and motor system toward a sensorimotor approach to joint attention and attempts to address a critical question: How does a deictic gesture affect an agent's perception?

In addition to contributing to human-robot interaction and socially inspired artificial intelligence, SHARE is supported by online contributors annotating and *indexing* the environment with the robot. Indexing is the maintenance of a set of hypotheses of recognizable parts of the environment. While other robotic architectures envision the robot operating alone or even sometimes with a social partner, this system acknowledges the constant access to social, on-line participants even when no one else is present. This provides an ever-present social partner even when the robot is seemingly alone. The NIMBUS architecture unifies long term memory

using a database to store all learning examples over the course of its lifespan, providing a substrate for long-term learning. NIMBUS also provides a caching system and pipeline for an embodied agent to capitalize on always-present, on-line participants. As long as the robot can connect to the Internet, it may query the Internet on-demand for new symbols

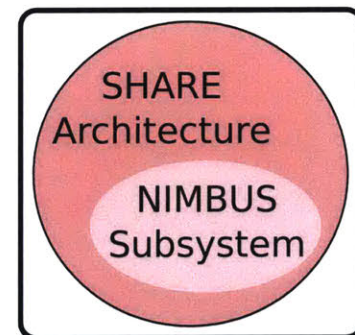


Figure 1-1: Situating the NIMBUS subsystem within the larger architecture (SHARE).

to map to its environment. NIMBUS provides an always-on server to take these queries, cache them, aggregate responses and respond to a robot’s on-demand queries. In the end, the robot uses a single point of entry (i.e., a single function call) to NIMBUS that operates both synchronously or asynchronously to provide lexical labels to the environment. NIMBUS handles query maintenance and long-term storage to map instances to labels and may aggregate these queries for learning at a later point. At the time of writing, the NIMBUS architecture has been running on an IBM Blade server for three years and has supported much of the work in this dissertation.

This dissertation contributes a number of empirical observations for systems inspired to:

- implement a deeper model of joint attention for robots,
- implement a synthetic model of visual attention, and
- contribute a human-robot interaction experiment in longitudinal joint attention and recall.

This work focuses primarily on the following factors:

- an analysis of eye gaze negotiation as the dynamic process of gaze following;
- an analysis of the exchange of deictic gesture directed toward the perception system;
- an analysis of memory recall between a human, a robot that directs it towards a stimulus and a dynamic scene.

## **1.1 Motivation**

For decades, psychologists (James, 1890; Seemann, 2011), neuropsychologists (Hebb, 1949) and more recently robotic architects (Ullman, 1996; Rao, 1998; Scassellati, 2001; Bridewell & Bello, 2015) have investigated the central mystery of how most animals have the ability to make sense of the “blooming, buzzing”(James, 1890) world around them. Many scholars believe our ability to cope with the morass of stimuli depends upon our

ability to inhibit, or ignore, peripheral and irrelevant stimuli and focus in on the most relevant features in that moment.

This dissertation is primarily motivated by the idea that this core coping mechanism impacts and relates to interactive joint attention. Additionally, this work is motivated by the idea that crowdsourcing offers an avenue to bootstrap interactions and experiments for a situated robotic architecture by collecting datasets quicker than ever before. In addition, this research is conducted within an interesting problem space: that of learning to name and index objects in the robot's environment through joint attention. The focus of this work in joint attention is primarily on object extraction, labeling and gaze manipulation.

## **1.2 Objectives**

This dissertation aims to understand and model an interaction between a human and a robot in the context of both static and dynamic environments. Referencing behavior plays a critical role in this interaction for both perceptual selection and directing the perception system of another agent. To engage with this problem, this work defines the following objectives:

- Define a model of perception that selects before it recognizes; that is, a model of visual search.
- Use a simple model of deictic referencing rooted in sensory selection that is compatible with this model of visual search.
- Understand how useful this model is in a human-robot interaction experiment.
- Define a subsystem within SHARE to aid in data annotation called NIMBUS.
- Investigate the behavioral dynamic of joint attention in dynamic scenes.

### **1.2.1 Stance and statement**

Attention directing in human-robot interaction involves situated perceptual systems and joint attention directly contributes to sustainable interaction and engagement.

## **Situated and embodied eye gaze negotiation**

Questions regarding artificially intelligent machines have long been focused on systems that can learn, adapt and act with new representations. Taking it further, attentional systems make it a clear objective to decide and maintain (i.e., update) a small subset of relevant features to the agent (Helgason, 2013) while still allowing the agent to transition to other features. The challenge for joint attention is to ensure that a social other can manipulate and shape this subset in meaningful ways. Only within the last couple of decades, researchers in the bio-inspired robotics field have begun to understand the importance of a social partner in shaping the relevant features that the robot should attend to during learning and referencing (Scassellati, 2001; Kaplan & Hafner, 2006; Nagai, Hosoda, Morita, & Asada, 2003; Ognibene & Baldassare, 2015; Steels & Hild, 2012). Unfortunately, little is known about how this socio-dynamical process proceeds over time in human-robot interaction. This research seeks to prioritize the design of robotic behavior to understand and address questions regarding the social exchange of gestural referencing and perceptual selection behavior in a human-robot experiment as the two interact within a dynamic (i.e., moving) shared scene.

This dissertation presents a synthetic attention system that may play critical roles in negotiating relevant and shared features between a human social partner and a robot. Inspired by behavioral and cognitive literature in joint attention (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998; Gallese & Goldman, 1998; Tomasello, 1995, 2000; A. M. Treisman & Gelade, 1980), the perception system is designed for synthetic visual search behavior and social negotiation (Chapter 4). This system is validated through real interaction with a human partner where the objective is to jointly select the same stimulus (Chapter 6). Clear contributions to joint attention build on many domains where feature sharing between a human and a robot is a primary concern, such as human-robot collaboration, rapid interactive robotic reprogramming and human robot dialogue. In each of these three domains, an underlying set of features must be chosen in which the robotic agent is to achieve its goals, learn new information and resolve word references, respectively. This shared interest in joint attention makes feature selection within the environment critical to

the functioning of robots in all of these domains. The success of each of these collaboration domains depends heavily on the state space the robot is currently programmed to consider.

Psychological and cognitive approaches to understanding eye behavior are roughly focused on characterizing two separate contributions to the set of features that make up the agent’s experience: *exogenous and endogenous factors* to the mind (Monsell & Driver, 2000). *Exogenous factors* can be seen from the computational view as external to the agent, having originated from the underlying sensor systems and made relevant to the agent’s state. One example of exogenous factors having a clear impact on human behavior is when something very loud happens in a room and everyone is directed toward this stimulus. Someone then asks “Did you **hear** *that*?” If a synthetic agent has not routed the underlying features in the proper way, the underlying stimulus is implicitly ignored by design. This has serious implications for language understanding. *Exophoric reference resolution* is the ability to resolve the word references to raw stimuli that were enumerated amongst the hypotheses. Therefore, the ambiguity present in the pronoun *that* is resolved. For exogenous factors, the relevant stimuli is asserted into the agent’s space of consideration. Likewise, *endogenous factors* assert themselves from deep inside the agent’s architecture and are not behaviorally visible but still have considerable impact on the agent’s outward behavior.

These two factors have also been called *bottom-up* factors and *top-down* factors when used with respect to gaze behavior (Pinto, van der Leij, Sligte, Lamme, & Scholte, 2013). Top-down approaches to gaze behavior operate primarily through an internal system that tightly controls sensing and action toward some goal. Bottom-up approaches tend to employ direct sensor-to-action or *reactive* centered models when generating gaze behavior. In the SHARE architecture presented throughout this dissertation (detailed most in Chapter 4), these bottom-up and top-down factors are unified by using a novel *simulation of other* mechanism to resolve competition between the two underlying modules. Any situated vision system that can take actions to direct its focus of attention must negotiate where to fixate next with respect to its goals. This particular system focuses on a socially inspired architecture that negotiates the next point of fixation with respect to its social partner. This system is validated with real interaction partners in Chapter 6.

## Driving gaze: the saliency system

Artificial saliency models, like those found by Itti and colleagues (Itti, Koch, & Niebur, 1998; Hou, Harel, & Koch, 2012; Harel, Koch, & Perona, 2006), have been successful in reproducing *saccade* behavior by reweighting filters (using a mechanism frequently called inhibiting return) in controlled ways to move the point of maximum saliency from location to location. A saccade is a movement of the eye from one target to the next. *Saliency* is defined in this context as a simple function that takes an image and returns a map of *conspicuity*, or readily noticeable fixation points. Factors like color, edge orientation and temporality of the feature are integrated in various ways to model this conspicuity.

The objective of such saliency systems is not to get the fixation order or trajectory through an image correct, but to locate points of fixation that resemble fixation locations that humans make on similar images. As a basic model of attention, saliency maps provide a map where all locations are available to higher order algorithms (e.g., that resolve linguistic utterances, like the example of a loud sound in the previous section). This has made saliency models a popular option for those studying attention in general (Itti et al., 1998; Tsotsos, 2011). Unfortunately, this approach doesn't provide a way to recognize or to simply index the features at the fixation location. For social saliency systems like those found in Marek Doniec or in Yukie Nagai's work (Doniec, Sun, & Scassellati, 2006; Nagai et al., 2003), the selection of salient positions becomes easier to resolve from a visual field alone, essentially allowing an embodied robot to share gaze direction. As mentioned before, saliency systems do not have the capability to recognize the object at the point of fixation. However, the system presented in this dissertation combines saliency with higher level recognition systems so the robot is not just fixated on a location but actually recognizes the features being referenced.

This work attempts to put these saliency systems in competition with a visual search mechanism (Chapter 4) during social interaction. The SHARE system unifies the ability to direct its gaze and to tether a symbol or **index** that same location. Psychology (specifically evolutionary psychology) has made apparent the clear social bias that occurs in human attention systems: attention is primarily directed to features that others attend to and the

features others are directing us to observe. This dissertation focuses primarily on agent-human-environment relations rather than other contexts where saliency systems may show up (i.e., agent-environment relations). This work incorporates not just the ability to behaviorally direct the robot's attention to something else in the environment but to also recognize the object. This type of attention enables a more situated, affordance-centered (robot-human-object-action relationships and thus predictions) theory-of-mind capability that could be leveraged in the future. Next, I will introduce a basic mechanism used in top-down attention architectures to continually re-task the attention system.

## **Goal-directed object search**

Significant questions still remain regarding how human partners direct and interpret gaze and attention. One of the key questions is how do pure saliency approaches lead to object recognition and goal-directed object search or activity recognition? Early work on gaze following has shown up in social robotics (Scassellati, 2001) and developmental psychology (Seemann, 2011). The only way researchers understood the process of social directing of gaze was as an endogenous process that involves recognition of gesture and goal-oriented object search. Recent systems integrate these problems into a single sequential context switching process (Ognibene, Chinellato, Sarabia, & Demiris, 2013). Clearly, the process must involve a much higher level of understanding than the modulation of simple saliency maps. Some researchers postulate that such goal-directed mechanisms constitute a set of primitives in which high level *visual routines* trigger sequential visual search goals (Ullman, 1996). Visual routines are simply higher level functions that search images to resolve queries such as “is object A inside object B?” This query can be resolved by fixating on A and searching around object A for boundaries of containment. Researchers have successfully implemented visual routines for games like Sonja (Chapman, 1990) and Pengi (Agre & Chapman, 1987). However to date, visual search as a substrate for visual routines has not proven successful for everyday images, though a scant few options have been proposed (Rao, 1998; Forbus, Mahoney, & Dill, 2002). The architecture presented in this document addresses two different issues at the heart of visual search:

1. foreground extraction and recognition in Chapter 4;
2. an effort to learn to map indices through reference following using real world images in Chapter 5.

While it is challenging to quantify, model and attribute the underlying factors that direct gaze, it is clear that these factors exist in biological systems (Monsell & Driver, 2000). Multiple researchers argue gaze direction factors should exist in human-robot interaction systems (Huang & Thomaz, 2011; Kaplan & Hafner, 2006; Nagai et al., 2003; Yu, Scheutz, & Schermerhorn, 2010; Nagai & Rohlfsing, 2009). For this reason, the SHARE architecture takes an embodied view of not just eye gaze as goal-directed action but also arm gestures as social referencing action directed toward a partner's biological attention system. The robotic architecture decides to move its own eyes by overriding the saliency model during goal-directed visual search and lets the gaze process "float" when visual search is not engaged by higher level processes.

## **Crowdsourcing and robotics**

Crowdsourcing is a new technique that leverages massive parallel user interaction to perform work in short periods of time. It can also source knowledge from novices and sometimes experts to perform real work effectively. Crowdsourcing has proven itself to be a promising avenue to rapidly bootstrap interactions and data from everyday people across domains. Crowds of people available on the Internet offer a level of parallelism that could potentially build large knowledge bases quickly and effectively. Computer systems that leverage crowdsourcing have shown progress in improving social skills through mutual criticism, collaborative writing and cracking challenging cyphers. This promising research has led many in machine learning and embodied robotics to use these systems to construct new and appropriate datasets quickly and effectively. Researchers have introduced crowdsourcing work in the synthetic agent literature, followed by crowdsourcing in robotics which deals with more situated problems.

Crowdsourcing itself is not a radically new idea. In fact, the idea of asking large crowds of people to perform work has been around for centuries. Grier (Grier, 2011) notes that



the idea may be traceable back to Charles Babbage (Babbage, 1832). Grier writes that editors even constructed the almanac of the 1820s through the coordinated work of many people using a process of assigning subtasks to qualified workers, awarding the work to the lowest bidder. While slower at that time, this coordinated and creative process is even more accessible and quicker today with the help of extremely fast communication networks like the present day Internet.

One of the earliest examples of collecting ongoing large datasets from on-line crowds was the OpenMind Initiative (Stork, 1999; Singh et al., 2002). The objective of the OpenMind Initiative was to source large datasets of logical statements that could be used toward reasoning in common sense ways. Following this early work, van Ahn popularized the idea that crowdsourcing can be an on-demand rather than an always-on process. For example, crowdsourcing can be used to focus on time-sensitive questions like cracking CAPTCHAs (Von Ahn, Maurer, McMillen, Abraham, & Blum, 2008). Crowdsourcing research has also led to experiments in motivating people to provide data through on-line entertainment like video games (Von Ahn & Dabbish, 2008; Orkin, 2013). Researchers use games to collect context-sensitive behavior performed in specific role-playing environments. These early experiments crowdsourcing context-sensitive behavior were performed within a video game called Mars Escape (DePalma, Chernova, & Breazeal, 2011; Breazeal, DePalma, Orkin, Chernova, & Jung, 2013). This work replicated the on-line environment using similar props to allow the features to map appropriately. Behaviors collected on-line were used in the virtual environment to train the robot and demonstrate multi-modal expression of the sourced behavior through memory-based autonomy (i.e., case-based reasoning). These robotic behaviors were sensitive to both task and environment but were not generalizable.

Crowdsourcing in robotics has primarily been defined as subsystem dependent, focusing primarily on vision, trajectory learning or allowing participants direct control of the robot. A considerable challenge to robotics and crowdsourcing is threefold:

- The interface design for each subsystem is context sensitive, meaning the design requires substantial work for it to be robust and usable by participants on-line as well as to ensure effective data delivery in real time.

- On-line participants typically provide data of poor quality. Researchers and engineers must frequently clean and check for data outliers from participants who either did not perform the task well or take it seriously.
- The crowdsourced dataset frequently requires a significant amount of human tending since the data collected in a batch could contain offensive actions or comments (Breazeal et al., 2013). Systems like Microsoft’s Tay are a good example of using data directly from all participants before being verified by consensus. Because the data is typically collected in batch and not as an ongoing and dynamic on-demand process, participants do not provide oversight or democratically police themselves.

Unlike these previous examples of crowdsourcing in robotics using batch collection, the NIMBUS subsystem presented in Chapter 2 provides an asynchronous system that provides both the features of on-demand data sourcing as well as ongoing batch data collection. The design and use of NIMBUS attempts to address the three concerns by designing each interface piece-by-piece for each ongoing, crowdsourced subsystem. Through the design and consideration of the expressiveness of the crowdsourced data, this architecture does not provide an avenue for offensive responses to be sourced from the crowd. In summary, the system presented primarily in Chapter 2 uses a principle of “vox populi” (Galton, 1907) (i.e., the “wisdom of the crowds”) with additional caching and redundancy to allow the system to adjust and monitor the data while it proceeds from participant to participant.

## 1.3 Contributions

The following chapters provide a list of questions, contributions and the set of computational models used in the experiments:

- After this first introductory chapter, Chapter 2 introduces the reader to the set of commitments and the agent architecture used throughout the experiments. While some chapters rely on a different set of group-wide commitments held by the research group, Chapter 2 introduces the reader to the pluggable architecture that complements previous work. Of special interest here is a commitment to long term memory

and stackable cached short term memory to support the asynchronous nature of cloud robotics.

- Chapter 3 introduces the first observation that foreground negotiation is a critical factor in coupling joint perception since humans have the capability to recognize objects in a gestalt way. This chapter provides the core argument for the following chapter: foregrounds offer a metric of success that allow behavior-centered observations and interaction scientists to ask the underlying recognition system (in a generative way) about the stimulus the robot is observing while allowing for self-report by participants. This metric may aid joint template construction toward situated visual learning.
- Chapter 4 provides an architecture that supports inquiry in social-perception learning using the metric described in Chapter 3. This chapter constructs a visual search algorithm that can recognize objects from fixation, incorporate human bias through an underlying saliency system and also put these factors in competition through a novel deictic simulation system. It describes the result of a simple experiment and shows that competition can resolve referential resolution conflict between the underlying subsystems.
- While Chapter 4 provides a blue-sky account of a system that can reason directly about the foreground, Chapter 5 capitalizes on already selected foregrounds to learn robust indexical references. The system presented in Chapter 5 uses a modified and novel recurrent neural network that learns saccade trajectories across a template to emit classification events. This system demonstrates that by biasing the learning algorithm with a deictic gesture signal, visual learning is accelerated.
- Chapter 6 wraps up by coupling visual saccade behavior with a human participant. In an attempt to guide the participant at certain moments of time, the robot uses deictic referencing to direct the human at critical moments in a dynamic visual narrative. The participants were probed for memory recall, measured for reaction latency when correcting the robot's gaze trajectory and tested to ensure they were following the

referential action.

- Finally, this dissertation concludes with the future direction of this work in Chapter 7.

# References

- Adalgeirsson, S. O. (2014). *Mind-theoretic planning for social robots* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Agre, P. E., & Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *AAAI Magazine* (Vol. 87, pp. 286–272).
- Babbage, C. (1832). *On the economy of machinery and manufactures*. Taylor & Francis.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(04), 723–742.
- Breazeal, C., DePalma, N., Orkin, J., Chernova, S., & Jung, M. (2013). Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1), 82–111.
- Bridewell, W., & Bello, P. (2015). Incremental object perception in an attention-driven cognitive architecture. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society* (pp. 279–284).
- Butterworth, G., & Cochran, E. (1980). Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development*, 3(3), 253–272.
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.
- Chapman, D. (1990). *Vision, instruction and action* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- DePalma, N., Chernova, S., & Breazeal, C. (2011). Leveraging online virtual agents to crowdsource human-robot interaction. In *Proceedings of CHI Workshop on Crowdsourcing and Human Computation*.

- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1), 193–222.
- Dix, A. (2003). *Human computer interaction* (Third ed.). Pearson Education Limited.
- Doniec, M. W., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *6th IEEE-RAS International Conference on Humanoid Robots* (pp. 34–39).
- Forbus, K. D., Mahoney, J. V., & Dill, K. (2002). How qualitative spatial reasoning can improve strategy game AIs. *IEEE Intelligent Systems*, 17(4), 25–30.
- Frintrop, S. (2006). *Vocus: A visual attention system for object detection and goal-directed search* (Vol. 3899). Springer.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in cognitive sciences*, 2(12), 493–501.
- Galton, F. (1907). Vox populi (the wisdom of crowds). *Nature*, 75(7), 450–451.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Psychology Press.
- Gray, J. V. (2010). *Reusing a robot's behavioral mechanisms to model and manipulate human mental states* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Grier, D. A. (2011). Foundational issues in human computing and crowdsourcing. In *Position Paper for the CHI Workshop on Crowdsourcing and Human Computation*.
- Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in Neural Information Processing Systems* (pp. 545–552).
- Hebb, D. (1949). *The organization of behavior; a neuropsychological theory*. Wiley.
- Helgason, H. P. (2013). *General attention mechanism for artificial intelligence systems* (Unpublished doctoral dissertation). Reykjavik University.
- Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1), 194–201.
- Huang, C.-M., & Thomaz, A. L. (2011). Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *IEEE International Conference on Robot and Human Interactive Communication* (pp. 65–71).
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,

20(11), 1254–1259.

James, W. (1890). *The principles of psychology*. Read Books Ltd.

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2), 135–169.

Monsell, S., & Driver, J. (2000). *Control of cognitive processes: Attention and performance* (Vol. 18). MIT Press.

Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, 15(4), 211–229.

Nagai, Y., & Rohlfsing, K. J. (2009). Computational analysis of motionese toward scaffolding robot action learning. *IEEE Transactions on Autonomous Mental Development*, 1(1), 44–54.

Ognibene, D., & Baldassare, G. (2015). Ecological active vision: Four bioinspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Transactions on Autonomous Mental Development*, 7(1), 3–25.

Ognibene, D., Chinellato, E., Sarabia, M., & Demiris, Y. (2013). Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspiration & biomimetics*, 8(3), 035002.

Orkin, J. D. (2013). *Collective artificial intelligence: simulated role-playing from crowd-sourced data* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Pinto, Y., van der Leij, A. R., Sligte, I. G., Lamme, V. A., & Scholte, H. S. (2013). Bottom-up and top-down attention are independent. *Journal of Vision*, 13(3), 16–16.

Rao, S. (1998). *Visual routines and attention* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Scassellati, B. (2001). *Foundations for a theory of mind for a humanoid robot* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Seemann, A. (2011). *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*. MIT Press.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open mind

- common sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems* (pp. 1223–1237).
- Steels, L., & Hild, M. (2012). *Language grounding in robots*. Springer Science & Business Media.
- Stork, D. G. (1999). The open mind initiative. *IEEE Expert Systems and Their Applications*.
- Tang, Y., Srivastava, N., & Salakhutdinov, R. R. (2014). Learning generative models with visual attention. In *Advances in neural information processing systems* (pp. 1808–1816).
- Tomasello, M. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, 103–130.
- Tomasello, M. (2000). *The cultural origins of human cognition*. Harvard University Press.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press.
- Ullman, S. (1996). *High-level vision: Object recognition and visual cognition* (Vol. 2). MIT Press. Cambridge, MA.
- Von Ahn, L., & Dabbish, L. (2008). Designing games with a purpose. *Communications of the ACM*, 51(8), 58–67.
- Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., & Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895), 1465–1468.
- Yu, C., Scheutz, M., & Schermerhorn, P. (2010). Investigating Multimodal Real-Time Patterns of Joint Attention in an HRI Word Learning Task. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 309–316).
- Yu, C., & Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*.



## Chapter 2

### Robotic architecture overview

The SHARE architecture is designed to better understand low-level interactions between perception and referencing in sensorimotor joint attention experiments within a crowd-present ecosystem. To be clear, SHARE is considered an architecture because it acknowledges the combined interactions between a number of underlying subsystems (sensing and action). It also requires caching to delay response from modular requests across two typically separate subsystems. The SHARE system leverages two mutually beneficial facets toward lifelong engagement and learning through aspects of both crowdsourcing and the interaction. The mutual synergy of crowdsourcing and the interaction provide two asynchronous dynamic loops between a social partner (both on-line and in-person) and an agent architecture that must reason about users in different ways.

Cognitive scientists building synthetic agent architectures are interested primarily in modeling hypotheses and theories about the deep and internal dynamics of an agent interacting with its world. They are particularly interested in adaptability, usually defined as the flexibility and performance of agents across multiple contexts. While other architectures like ACT-R/E (Trafton et al., 2013) and Subsumption (R. Brooks, 1986) focus primarily on the internal dynamics toward generating robotic behavior, SHARE takes into account both biological inspiration and the synthetic nature of an Internet-connected, robotic architecture. By leveraging these two dynamic loops, SHARE can accelerate growth and learn faster through multiple interactions in parallel. This idea is demonstrated in Section 2.5, within simple symbol-to-sensor learning and through long-term interaction. The following

two sections document the architecture built to support the ongoing, three-year perceptual learning of this agent.

## 2.1 Features of SHARE

The following attributes of SHARE help define its unique performance capabilities. The SHARE architecture begins with modeling gaze through saliency hill jumping and a novel spotlight model of sensor-to-symbol mapping, allowing the robot to saccade to novel stimuli while also tethering a symbol to the stimuli. SHARE subsequently believes it recognizes the symbol underneath the point of fixation. This model of symbol tethering is closely related to and inspired by the FINST model of spatial indexing (Pylyshyn, 1989). To source the symbols, the architecture can recall symbols from memory that are stored in a database sourced on-line. This modular memory system operates across contexts both on-line and in real-world situations. SHARE's memory is also compatible with both the system on-line and the system operating in real-time on the robot. Additionally, the architecture moves from fixation to recognition behavior to full trajectory learning (i.e., fixation-to-fixation) which builds plausible gazing behavior.

The NIMBUS subsystem<sup>1</sup> is designed as an internal subsystem with a single entry point to the overall architecture and supports both synchronous and asynchronous crowd-querying. The NIMBUS subsystem provides the substrate and capability to interactively query participants on the Internet as well as the ability to index, reference, extract and learn directly from sensor data. The SHARE subcomponent provides the capability to interact with participants through simple referencing and extracting behavior. Together, these systems make up the agent architecture presented in this dissertation.

Crowdsourced robotic architectures that incorporate cloud-based resources are just now gaining popularity (Thielman, 2015). However, researchers have launched very few investigations into these capabilities which could support further claims past the point of speculation. In fact, some experiments have shown that pure and direct crowdsourcing can be dangerous. For example, on-line participants unwittingly trained Microsoft's Tay

---

<sup>1</sup>This resource is always online at <http://nimbus.media.mit.edu>.

to be racist through their questionable input. These observations are previously noted in past work (Breazeal, DePalma, Orkin, Chernova, & Jung, 2013). Fortunately, the NIMBUS design ensures that the crowdsourced data is cleaned through both redundancy and verification.

The rest of this chapter describes the architecture that is designed to support symbol-to-image mapping while ensuring quality results. This chapter covers the following:

- Section 2.2 discusses the contributions made throughout the document to human-robot interaction. This section provides a detailed account of the careful considerations required prior to performing research in joint attention for communication between humans and robots.
- Section 2.4 presents a system that supports synchronous and asynchronous crowdsourcing requests across a cloud-centered infrastructure.
- Section 2.5 presents a novel mechanism to improve quality consistently across all crowd requests.
- Finally, Section 2.6 describes how these subsystems integrate into the larger SHARE architecture to support the studies performed in Chapters 4, 5 and 6.

## **2.2 Loop 1: Joint attention for situated robotic interaction**

The rest of this dissertation will focus on the problem of *shared attention* between a human and a robot. While the previous chapter described the vocabulary and concepts needed to understand the problem, this section will focus on the various approaches that were considered prior to the development of this system. In this case, joint attention is a shared, synthetic attention system interacting with a human which is an incredibly challenging endeavor. This section will discuss the technical challenges that influence the design to successfully achieve shared attention.

Shared attention is one of the most important skills for an agent to acquire prior to language acquisition and comprehension. It is also critical for sustaining a human-robot

interaction (as seen in Chapter 6). The phenomena is most often discussed in the context of language learning; however, it is a significant skill for robots to possess since it shows up in many other domains. This fundamental problem is critical for robotic programming by demonstration in which features are shared prior to goal learning, for linguistic and exophoric referencing for human utterance understanding and, in general, for coordinated communication and sustained engagement between humans and robots. Chapters 4, 5, and 6 explore two major problems that impede progress in achieving sustained shared attention: 1) visual attention and 2) nonverbal gesture and gaze estimation. This chapter will focus primarily on *Loop 2*, the underlying architecture that supports the systems designed in the subsequent chapters.

Visual attention is a problematic aspect of joint attention that has received only limited examination in the computer vision community. The problem lies in recognition: given a fixation point, what is the agent choosing to recognize? In later sections, this problem is called *the shared visual field problem*. In general, when a human directs the gaze of the robot to attend to a particular stimulus and the robot is successfully directed through either an action driven by a learning algorithm or a behavior-based approach, a perception problem emerges. The robot is provided with new scene stimulus that it must assimilate and understand. There are two different approaches to incorporating recognition into this problem. The first approach is to perform semantic segmentation where the scene is segmented into components. For each component, recognition is performed across the entire scene in a convolutional way (left to right, top to bottom). Using this approach, the robot can filter out recognized objects that are extraneous. This approach may not be efficient since participants assume an agent is recognizing an object during fixation after the agent is directed toward the new target object. An alternate approach is to direct the agent toward a new fixation location, at which point recognition operates on just the point of fixation. A basic algorithm for this approach is demonstrated in Chapter 4 and extended in Chapter 5. Significantly more work is needed before a robot will have the capabilities of directed perception.

Nonverbal joint attention is typically cued from two different nonverbal signals: gaze direction and deictic gesture (Yu & Smith, 2016). Estimation and detection of these two

signals is key to the challenge of shared attention. The decision to direct the gaze of the robot is typically driven by the use of nonverbal gestures and verbal commands while gaze serves as a relatively weak signal to follow. Progress in these domains is limited to sensor and machine learning technologies that detect and estimate these signals when needed. The challenge inherent in detecting and estimating signals is often the amount of error captured in the signal. The best gaze estimation algorithms at the time of writing have upwards of  $20^\circ$  of error from ground truth (Morimoto & Mimica, 2005) while deictic gesture has similar estimation error profiles. State-of-the-art systems require proximal interactions (hand and eye trackers in close proximity of the user) before the error is significantly reduced. For this reason, much of the work in this document exists in proximal settings. Solving the problem of distal nonverbal directing and distal gaze estimation may require the development of other methods. Some ideas in this direction capitalize on interaction memory and other contextual clues as a means to resolve references. Unfortunately, there is a lack of good research in algorithmic artificial intelligence to define context-inspired solutions. Alternately, the system may benefit from enhanced methods and hardware to estimate distal-target gaze with precision.

This dissertation provides contributions to the problem of human-robot joint attention. Many of these contributions were only possible using crowdsourced datasets prior to the interaction. The rest of this study will detail these contributions but next, here is an overview of the architecture developed to support this research.

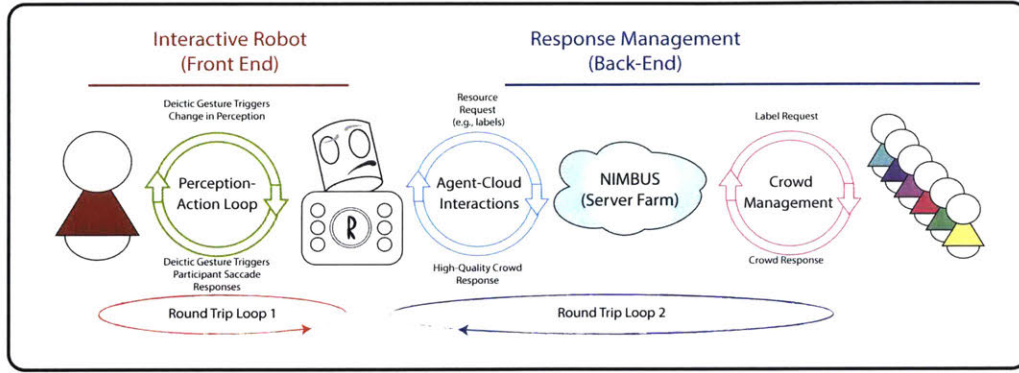
## **2.3 Loop 2: On-the-fly crowdsourcing infrastructure**

Few socially intelligent robots support integrated real-time crowdsourcing and cloud-based resource management into their robotic architecture. Many surveys cover the broad spectrum of research that utilizes crowdsourcing in robotic learning tasks, ranging from perception learning (Russell, Torralba, Murphy, & Freeman, 2008) to action learning (Sorokin, Berenson, Srinivasa, & Hebert, 2010). For a survey of various projects that use crowdsourcing in machine learning and robotic tasks, the comprehensive reviews by Goldberg and Kehoe et. al. (Goldberg & Kehoe, 2013; Kehoe, Patil, Abbeel, & Goldberg, 2015)

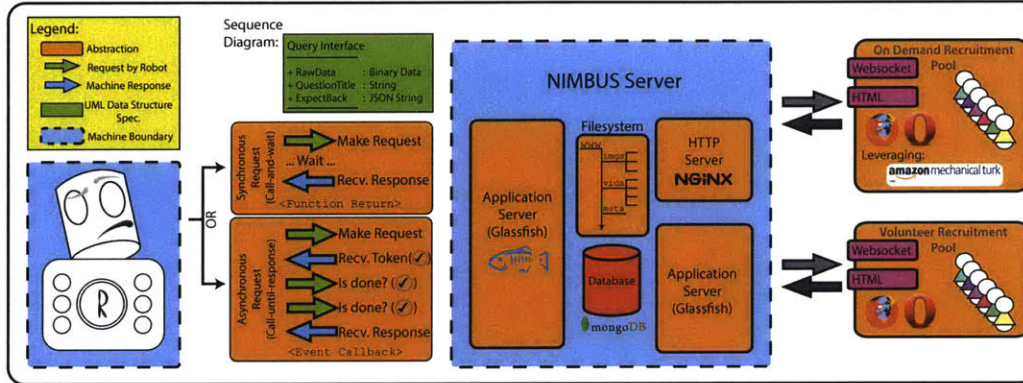
offer a detailed landscape. As an example, the RoboEarth project vision is still the most mature vision of cloud resources integrating back-end architectures to support perception, action and mapping (Molengraft, 2016). Rapyuta, the underlying framework, can be very powerful but its emphasis on non-interactive tasks makes real-time latency issues secondary for their level of inquiry (Hunziker, Gajamohan, Waibel, & D'Andrea, 2013). Details on many of these architectures are sparse, with Rapyuta (Hunziker et al., 2013) and the work by Osentoski et. al. utilizing closed-loop control (Osentoski, Crick, Jay, & Jenkins, 2010) with a single on-line participant providing the most concrete detail. Other architectures treat crowdsourcing as a secondary or tertiary characteristic of the research process, performed as data collection or after data collection to support annotation.

Previous research does not utilize the cloud as an alternative resource to employ autonomously when other resources are unavailable. Rapyuta manages to schedule available cloud resources for the overall real-time system using specialized scheduling algorithms. Large tasks can be scheduled to operate on cloud-based hardware or on local embedded systems inside of the robot, depending on the needs and demands of the entire architecture. This newly conceived system focuses instead on the labeling task as a source of information that can be made available on-demand through either a social interaction between the interaction partner and the robot or as an interaction the robot can request on-line. The concept of managing cloud-based resources through scheduling is very important and is complementary to the architecture designed and documented in this chapter. Systems and architecture pipelines to support robotic interaction with either the environment or with a social other are at present not a design concern for contemporary robotic architectures.

In this work, a crowdsourcing system is used to support interactive perceptual learning between the robot and its environment. To support this level of inquiry, this work reviews real-time perceptual learning systems. Raptor (Goehring, Hoffman, Rodner, Saenko, & Darrell, 2014) is a recent (and closely related) successful dynamic training process that utilizes the power of ImageNet (Deng et al., 2009), a large database of object categories. With ImageNet, Raptor can discover early image priors for in-situ training purposes. Researchers use ImageNet exclusively for figure-ground patch discovery and employ these image patches to dynamically train fast, whitened, HOG-based object detectors during



(a)



(b)

Figure 2-1: a) Long term vision of longitudinal embodied visual learning, b) system architecture and design of the NIMBUS architecture: sequence diagram, server subsystem and caching mechanisms.

an interaction. This process takes under 30 seconds for a new image patch to reify into an object detector. The perception-action loop extends this type of system and allows the robot to take deictic actions toward resolving figure-ground ambiguity to support object discovery with a social partner. The following section supports this perception-action loop documented more thoroughly in Chapter 4 while focusing primarily on incorporating crowd-based labels into this cognitive architecture.

## 2.4 Architecture layout and pipeline

The SHARE system is designed around a single principle: when the agent doesn't know a word for an object it is observing, the agent asks a social partner. Then the robot can

either share attention with a social partner who is intimately present with the robot using joint attention or it can extract the scene and ask the users on-line to structure the region and symbol to be mapped to that region. Designers of a system must ask when the robotic agent understands it is observing an object not in its database. I address this question and provide a solution in Chapter 4.

Figure 2-1a visualizes the basic architecture while Figure 2-1b visualizes the underlying caching system that supported the architecture. First, the agent architecture has the ability to saccade around its environment and recognize elements that are already mapped. This piece of the architecture is described in detail in Chapter 4 (and also extended in Chapter 5). These chapters explore how the agent finds, extracts and learns about its environment in a basic setting. When the agent is signaled internally that it doesn't recognize the stimulus, a cascade is triggered in its architecture to crowdsource the answer. The cognitive architecture computes a signal through a competition in which the stimulus wins against the recognition algorithm for a mapping that the architecture perceives.

Figure 2-1b describes the subsystem that must crowdsource an answer for the mapping region to symbol. The system slightly resembles the LabelMe project (Russell et al., 2008), in which website visitors are provided images and the visitors must lasso-select the given region and provide a label to the symbol. The difference between NIMBUS and LabelMe is twofold. First, NIMBUS is designed to be operated on-the-fly. It attempts to crowdsource the answer quickly and effectively, rather than acting as an ongoing system with no temporal constraints. Second, NIMBUS provides an interface to an always-on system that the robot may join and use. Figure 2-1b walks the reader through this process.

First, the robot securely connects to the server using an authorized, unique token. Once connected, the robot can make a number of requests, like requesting a label to an image or requesting a region to be selected. It may also query a large database for images, labels or past interactions with participants for later analysis. Each visiting participant interacting with the system is identified with a cookie containing a simple global unique identifier (GUID) or with their unique Mechanical Turk worker ID. The system can push both HTML and Javascript to the visitor's experience as well as a supporting application server. These systems have access at their API level to store data in the MongoDB through simple re-



quests to the webserver and application server. The application server provides a simple way to also store other data types (like video content) that are not well-suited for storage in the MongoDB. In addition, the robot's API has the capability to ask sweeping statistical queries across the database through the use of secure read-eval-print (REPL) loop connections. Figure 2-1b, the basic NIMBUS subsystem, provides a system for always-on and always-learning interactions for the robot and has been in operation for approximately three years.

The NIMBUS architecture was integrated into a previous robotic architecture that was built over the past few decades around the idea of synthetic lifelike agents (Burke, Isla, Downie, Ivanov, & Blumberg, 2001). NIMBUS architecture was later adapted into a more embodied approach that takes into account sensors and motors. It incorporates elements of interactive task learning (Thomaz, 2006) and principles of animation (Breazeal et al., 2003). Recently, the architecture has forked into two separate initiatives: one that emphasizes the ability to take advantage of small, embedded and ever improving smart-phones (K. Williams & Breazeal, 2013; K. J. Williams, Peters, & Breazeal, 2013; Gordon, Breazeal, & Engel, 2015) and another that emphasizes its ability to operate across small and large scale computer systems (DePalma, Chernova, & Breazeal, 2011; Breazeal et al., 2013). This section will focus on a simple pipeline to unify these movements and support lightweight requests for resources not available to the robot alone. With this innovation, the robot can make decisions to connect with individuals on the Internet, based on the availability of resources and participants not present at the time of interaction.

Next, let's explore a core mechanism used across the architecture to improve the quality of results sourced on-line as documented in Section 2.6.

## **2.5 A core mechanism to improve contribution quality**

This section proposes a novel method to exchange quality of response for speed of response from people on the Internet. Further, this proposal is backed with empirical findings from experiments performed with the Amazon Mechanical Turk. This method shows an improvement in quality as an alternative to response time using this robotic architecture. This

novel mechanism is embedded in the NIMBUS architecture, as described in Section 2.4. In isolation, NIMBUS allows us to perform research for architectures that take into account crowdsourcing as a core part of their interactions. NIMBUS serves to collect and coordinate information from an always-on and resilient data collection interface, to crowdsource on-demand when needed by the robot and to act as a computational resource for highly intensive processes.

The trade-off between the quality of response from people on the Internet and the speed at which a response can be obtained by an autonomous system is still not well understood. The work contained in this document traces the past few years of engineering toward building a resilient cloud infrastructure to support robots on the ground and in the field. One of the higher level features of the system is to retrieve information from databases in the cloud, and when unavailable, to obtain high quality labels from crowds of people on systems such as Amazon Mechanical Turk or Crowd Flower. This research measures round trip time from the robot to the crowd and back while discussing expectations for a system that leverages crowdsourcing as a core cognitive resource. Next, this chapter introduces a novel mechanism to improve quality of the response at the expense of response speed. This chapter concludes with remarks and thoughts about the future of technology built for the purpose of efficient and interactive labeling to support interaction.

### **2.5.1 Problem domain: on-the-fly visual labeling**

This system seeks to realize the simple abstraction that perception-action loops within interaction help define an interactive and situated visual labeling task and that back-end loops completed between the real-time embedded system and the servers support the needs of the interactive system (see Figure 2-1). While previous research in crowdsourcing and robotics focuses on collecting datasets and direct interaction for behavior control, this research is based on a visual labeling task. While behavior in interaction can be highly dependent on context, visual learning and indexing requires sizable datasets and clearly can be reused over long periods of time. Interactive visual labeling is still a large, open problem space rich with potential breakthroughs. Using it, a system can take early visual stimuli and dynam-

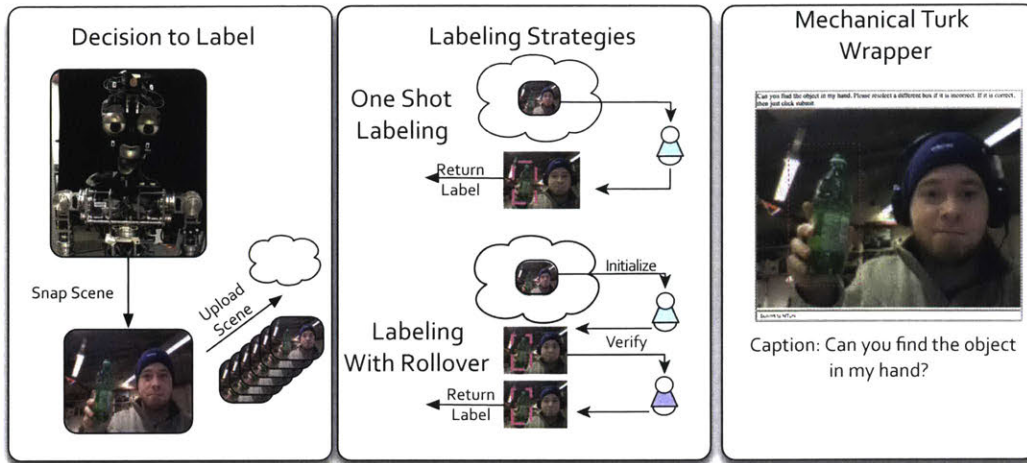


Figure 2-2: Data collection, conditions and user interface for crowdsourcing labels for a situated robotic agent.

ically create stable object detectors or abstract deictic pointers (A. G. Brooks & Breazeal, 2006; Ballard, Hayhoe, Pook, & Rao, 1997) in real time.

To demonstrate the power of this construct, this system is built to learn labels through both interaction and a server architecture that manages an Amazon Mechanical Turk account. Figure 2-1 shows the cognitive architecture that surrounds the system. At the interaction level, the system utilizes gaze and deictic gesture to visually guide the robot toward relevant areas in the scene. It uses finite horizon reasoning and simple low-level cues to visually extract select image patch foregrounds from the scene. This interactive guiding process allows the robot to be directed toward stimuli it normally detects if operated in a purely discretized, object-oriented manner. Additionally, it allows the robot to use those early stimuli as training instances for socially relevant deictic and lexical referencing. In the event the user disappears, the robot still has the ability to saccade to new stimuli and collect labels from Amazon Mechanical Turk for a small fee. The following two sections (Sections 2.5.3 and 2.5.4) introduce a natural tension within the architecture between speed of response and quality. This research introduces two methods that document the trade-off one could expect to see between systems of these types: *one-shot labeling* and *rollover labeling*. In one-shot labeling, the system attempts to collect the label as fast as possible. The other method, called rollover labeling, attempts to reliably secure a label in the presence of noisy, error-prone crowd responses by rolling over the result from one crowdsourcing

participant to another in an attempt to correct and clarify previous labels.

To test the performance of the proposed method, the following hypotheses were defined and tested with findings presented in the following sections:

- H1: Allowing participants to validate one another improves the reliability and quality of the result.
- H2: Due to the extra answer validation, the process of labeling with rollover will take more time.

### **2.5.2 Data collection**

To test the previous hypotheses, a camera was set up to snap a single image from its view-point and to push the image along the architecture documented in Figure 2-1b. For this simple test, the captured image was pipelined to the web server and wrapped in HTML. NIMBUS hosted the page on Amazon Mechanical Turk and used the jsPsych (De Leeuw, 2015) package to save the data back to Mechanical Turk as XML. NIMBUS probed Mechanical Turk every 1 second to check for an answer (see Figure 2-1b for a sequence diagram of asynchronous monitoring). If the answer was waiting, then the architecture auto-approved the data and extracted the XML data to use in a rollover. The process looks similar to the following itemized steps:

1. The camera takes a snapshot from the sensors of the robot's architecture.
2. Upload the data to the web server from the robot and wrap it in a visually understandable and descriptive manner.
3. Find an available on-line participant to label it, for a price. <sup>2</sup>
4. Wait for the participant to accept the job and read the job instructions.
5. Await the participant's response.
6. Optional experimental condition: extract response, re-wrap and repeat three times.

---

<sup>2</sup>This seems to be the slowest process.

The optional experimental condition used the current result from Mechanical Turk to embed the answer as a visual overlay on top of the query (see Figure 2-2 right, pink box). The answer was posted as an HTML GET request to the webpage and could be adjusted or corrected by the participant on-line to return a hypothetically more correct answer. To test elapsed time and quality, the system recorded elapsed time for one-shot and rollover requests of the crowd. Each condition (one-shot and rollover) was performed 20 times in sequence and, due to the fact that the system used multiple assignments within a HIT, ensured that each of 20 unique workers received 20 unique assignments. The baseline was computed globally for the image across both conditions using the mean of each bounding box vertex. Mean squared error and resulting variance were computed using the different set of pixels between the baseline and resulting bounding box.

### 2.5.3 Results: Time-to-response measurements

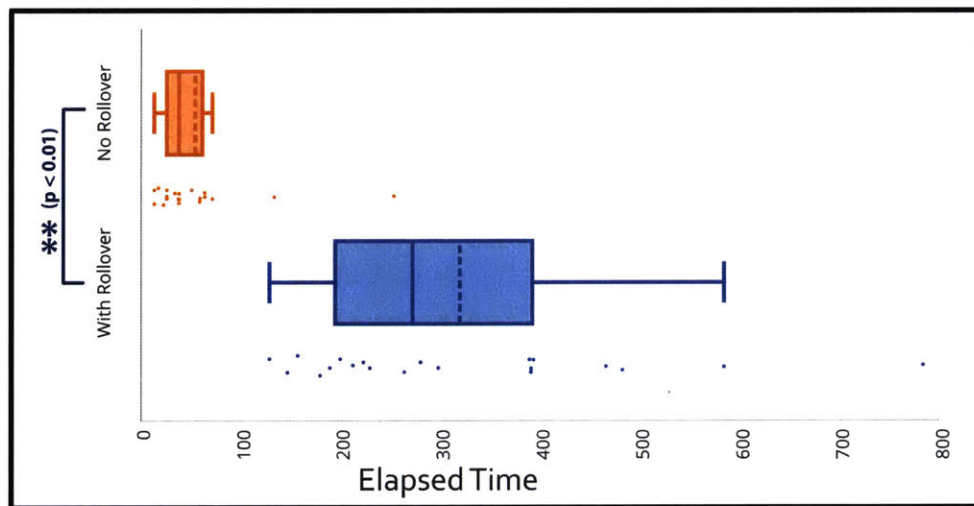


Figure 2-3: Elapsed times of the two conditions: with and without the novel rollover mechanism. Response times are significantly different ( $p < 0.01$ ) with one-shot labels outperforming rollover.

The image and requisite interface details were uploaded to NIMBUS and a process began to synchronously wait for a response. Total round-trip times are reported in Figure 2-3. This process used Amazon Mechanical Turk at a competitive rate of \$0.25 for a 45-second labeling task. The experiment was performed 20 times for each condition. Findings

show one-shot labeling can be a relatively quick process (Figure 2-3) while rollover can take a substantially longer length of time. Response times for one-shot labeling fell in the range of 12.9s, 520s with a median of 37.5s while a rollover process fell in the range of 95s, 1217.8s with a median of 263.58 seconds, or approximately 4.3 minutes. Response times were significantly different between the two conditions ( $p < 0.01$ ). It is clear that one-shot labeling has the potential to be a very fast process, optimistically only taking 13 seconds. Section 2.5.1 describes the vision of on-the-fly labeling and aiding robots through crowdsourcing. For this vision to come to fruition, researchers and engineers need to reduce response times substantially and while 13 seconds is optimistic, 30 seconds is typical for one-shot labeling. This finding suggests that asynchronous requests for labels and help from the on-line crowd should either a) make non-critical requests from the crowd or b) attempt better predictions of the future 30 seconds prior to crowdsourcing the query. Now let's explore the response itself to analyze the quality of response from the two conditions.

#### **2.5.4 Results: Analysis of quality and speed**

Figure 2-3 finds that receiving a response can be as quick, but what kind of data quality can one expect from a quick 30-second, one-shot label? In previous experiments (DePalma et al., 2011), it was found that the quality of on-line data can vary widely. Using the same results acquired from the 20-participant on-line experiment, the baseline answer is derived from calculating the mean bounding box. Following this reasoning, this study looks primarily at the resulting sum error across the condition, assuming the user of such a system will ask for a response once and expect a high quality answer as quickly as possible.

To measure the performance of these two conditions, this study calculates the mean squared error (MSE) with respect to the baseline and finds that rollover significantly reduced the overall error. Table 2.1 shows that while the answers can begin with many errors, rollover helps correct them. The resulting effect is that the MSE of the rollover strategy is significantly better than the one-shot labeling. Row 1 reports the errors from the two strategies. Row 2 reports the standard deviation of the answer. Interestingly, the error is minimized in the rollover condition (341.5 pixels vs. 518.6 pixels) by a factor of 1.5. Addi-

tionally, the responses from rollover provide a reduction in response variance (18.4 pixels vs. 22.73 pixels). This suggests that the contributors' edits to another's contribution provide a higher fidelity bounding box around the training instance and thus a higher quality result.

In the course of this experiment, other strategies also emerged as being salient. For example, while some of the bounding boxes and labels provided are rife with error, on average, Amazon Mechanical Turk workers provided accurate responses. The assumption (H1) was one-shot labeling will not provide an accurate result on average and rollover is necessary for accuracy. Instead, research found that with many answers, a researcher can statistically remove outliers based on the results of many parallel queries. This opens up the possibility to an alternate strategy from strict rollover to parallel one-shot queries that may allow the agent to rely on the consensus of many to remove the untrustworthiness of smaller numbers of individuals. Then the agent can simultaneously model these workers and weight them appropriately. This parallel process may simply capitalize on one-shot labeling speed and still keep high quality results.

Using these findings, these mechanisms are integrated into a broader architecture in the next section.

	One-Shot Labeling	With Rollover (Final Iteration)
MSE from baseline	518.6px	341.5px
Standard Deviation	22.73px	18.4px

Table 2.1: Quality of response between one-shot labeling and labeling with rollover reported as mean squared error and standard deviation with respect to baseline. Rollover outperforms one-shot labeling in terms of accuracy of response.

## 2.6 Extended cloud architecture

Utilizing the previous findings in Sections 2.5.3 and 2.5.4, an overall system was developed called NIMBUS to support crowdsourcing in future parts of this dissertation. While the system itself is not a core aspect of this research, much of the research demonstrated was impossible without this system. The NIMBUS system presented in Figure 2-4a describes

the layout of algorithms meant to improve themselves over time through interactions with the crowd. The system itself uses seven main parameters:

- A description of the experiment,
- A maximum number of participants,
- A total budget,
- A maximum elapsed time,
- Participants per iteration of the algorithm,
- The quality assurance function (see previous section),
- And, the maximum number of iterations.

Unfortunately, the latency of response found in the previous sections demonstrates the architecture primarily supports back-channelled, asynchronous requests. This allows the robot to continue the interaction rather than synchronously waiting for responses that may break an interaction. The architecture presented ensures participants are exposed to the experiment either within subject experiments or between subject experiments, depending on the configuration of the parameters. The system has access to an internal database that maintains running experiments as well as the set of long-term learned models (see Figure 2-4b, 2-4c and Figure 2-1). This database and the architecture used extensively for data collection are detailed in Chapter 4 and Chapter 5.

To understand the extent of the architecture pipeline, it's helpful to take a data collection request through the architecture. First, a simple API call triggers an experiment deployment into either production or sandbox as long as it conforms with the following interface:

- The experiment can render a stimulus as HTML viewable by a web-browser.
- It handles a websocket callback.
- It can provide a function that can summarize a small number of responses to a stimulus.



- It can provide an HTML renderable version of a consent form.
- It can provide a simple summary statistic that improves quality.
- The experiment provides a set of overall experimental parameters like maximum subjects, maximum budget, maximum time spent, participants per iteration and a learning algorithm that takes a previous state and a label as input and evolves the model one step for the next set of participants.
- The system iterates on the model until some termination condition, which is by default defined as exceeding a budget, exceeding time spent or exhausting a number of subjects.

Following setup, the system switches to production mode and has the ability to automatically monitor other global termination conditions like lifelong budget. Every experiment is tested extensively in a sandbox prior to production.

Once an experiment is triggered, it goes through a simple process of unraveling. First, the set of conditions are unrolled and an empty set of unique visitors are initialized. Each condition deploys its own set of assignments to Mechanical Turk of size representing the participants per iteration (*ppi*) of the algorithms evolution. Next the system renders a summary version of the experiment to draw in the participants. Once a participant accepts and completes the assignment, the response is recorded in the long-term database. The event-based system calls the *ppi* aggregator function once the number of assignments are completed and stored in the database. These responses are then aggregated and passed to the summary and quality assurance function to be cleaned and recorded in a separate record. The last model state is passed, along with the response by the participants, to an arbitrary learning algorithm to iterate once again. This process iterates, batch-after-batch, until the budget, time or number of participants is exceeded.

## **2.7 Designing an interactive perception-action system for shared visual attention: next steps**

The robotic architecture used throughout this chapter incorporates interactive perception-action loops and interactions between the robot and crowds of people on the Internet to learn to map words to its perception system. Section 2.5.3 demonstrates a clear need to reason about the speed of a response and the quality of a response a robot designer can expect. Even in the best circumstance, one-shot labeling can take at least 12.74 seconds and take upwards of 4 minutes, depending on the availability of participants on-line. Timing and scheduling the interaction options is an interesting dynamic challenge for this always-on hybrid architecture but progress is definitely possible. Following this basic architecture, this research presents the elements of SHARE that make up the core focus of this system (described in Section 2.2 as “Loop 1”): illustrating the synthetic and robotic visual attention system that incorporates deictic gesture following and recognition as well as sustaining this interaction over long periods of time.





# References

- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(04), 723–742.
- Breazeal, C., Brooks, A., Gray, J., Hancher, M., McBean, J., Stiehl, D., & Strickon, J. (2003). Interactive robot theatre. *Communications of the ACM*, 46(7), 76–85.
- Breazeal, C., DePalma, N., Orkin, J., Chernova, S., & Jung, M. (2013). Crowdsourcing human-robot interaction: New methods and system evaluation in a public environment. *Journal of Human-Robot Interaction*, 2(1), 82–111.
- Brooks, A. G., & Breazeal, C. (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (pp. 297–304).
- Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1), 14–23.
- Burke, R., Isla, D., Downie, M., Ivanov, Y., & Blumberg, B. (2001). *Creature smarts: The art and architecture of a virtual brain*.
- De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1), 1–12.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).
- DePalma, N., Chernova, S., & Breazeal, C. (2011). Leveraging online virtual agents to crowdsource human-robot interaction. In *Proceedings of CHI Workshop on Crowdsourcing and Human Computation*.
- Goehring, D., Hoffman, J., Rodner, E., Saenko, K., & Darrell, T. (2014). Interactive adap-

- tation of real-time object detectors. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1282–1289).
- Goldberg, K., & Kehoe, B. (2013). Cloud robotics and automation: A survey of related work. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2013-5*.
- Gordon, G., Breazeal, C., & Engel, S. (2015). Can children catch curiosity from a social robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 91–98).
- Hunziker, D., Gajamohan, M., Waibel, M., & D’Andrea, R. (2013). Rapyuta: The roboearth cloud engine. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 438–444).
- Kehoe, B., Patil, S., Abbeel, P., & Goldberg, K. (2015). A survey of research on cloud robotics and automation. *IEEE Transactions on Automation Science and Engineering*, 12(2), 398–409.
- Molengraft, M. (2016). *Roboearth — a world wide web for robots*. <http://www.roboearth.org>. RoboEarth, E.U. (Accessed January 2016)
- Morimoto, C., & Mimica, M. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1), 4–24.
- Osentoski, S., Crick, C., Jay, G., & Jenkins, O. C. (2010). Crowdsourcing for closed loop control. In *Proceedings of the NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.
- Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the first spatial-index model. *Cognition*, 32(1), 65–97.
- Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3), 157–173.
- Sorokin, A., Berenson, D., Srinivasa, S. S., & Hebert, M. (2010). People helping robots helping people: Crowdsourcing for grasping novel objects. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 2117–2122).
- Thielman, S. (2015). *Man behind darpa’s robotics challenge: robots will soon learn from*

*each other.* (The Guardian. Accessed January 2016)

- Thomaz, A. L. (2006). *Socially guided machine learning* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Trafton, G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., & Schultz, A. (2013). Act-r/e: An embodied cognitive architecture for human-robot interaction. *Journal of Human-Robot Interaction*, 2(1), 30–55.
- Williams, K., & Breazeal, C. (2013). Reducing driver task load and promoting sociability through an affective intelligent driving agent (AIDA). In *Human-Computer Interaction–INTERACT* (pp. 619–626). Springer.
- Williams, K. J., Peters, J. C., & Breazeal, C. L. (2013). Towards leveraging the driver's mobile device for an intelligent, sociable in-car robotic assistant. In *IEEE Conference on Intelligent Vehicles Symposium (IV)* (pp. 369–376).
- Yu, C., & Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*.





## **Chapter 3**

# **Human-robot shared attention : metric and analysis**

Human-robot attention sharing is linked to critical communicative competence for situated learning systems (Tomasello, 2000) and is considered to be a critical factor that differentiates humans from chimpanzees, (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998) opening the door to cultural, social learning. Attention sharing is also linked to triadic interactions between two situated agents and a present object, its affordances and relevant shared context in general (Trevarthen, 1979). Following this line of reasoning, if a robot can tap into this critical skill, it may tap into relevant social and cultural learning. Though many researchers and roboticists have focused on building models of joint attention to support socio-cultural learning systems for robots, there is still significant progress to come.

Human-robot interaction systems are engineered to study a) the effects of robotic action on an interaction partner and b) the effects of human action on a robotic agent architecture. This dissertation and document focuses primarily on a newly developed system (called SHARE) which uses shared attention as a mechanism that is useful for feature sharing (on the robot side). Chapter 1 focuses on an example that many scientists use in which simple utterances like “Help!” require context and dialogic interaction to sufficiently resolve. In the artificial intelligence field, the set of features, cases or frames needed to resolve such utterances is typically considered a sufficient solution. Shared attention systems argue that

the recall and selection of such features, cases or frames must be socially shaped and negotiated whether they are internal cases recalled or external features observed. This dissertation focuses on a visuo-social, situated (perceptual) cognitive system to support a higher level agent architecture (c6 (Burke, Isla, Downie, Ivanov, & Blumberg, 2001)), mainly due to the relevancy of this problem to situated robotics rather than more unembodied agents.

Unfortunately, there are a number of incredibly important problems that are *unsolved and unaddressed* in the literature that are absolutely critical to developing a synthetic shared attention system between a human and robot. For a situated visuo-motor system to support higher level cognition, many argue that synthetic visual attention must have a plausible way forward. Breaking these concerns down, this chapter argues that the following features must exist in a unified way in the architecture. Each feature includes an example of why it is important.

1. Visual search: this visual subsystem provides the capability to perform goal-directed exophoric reference resolution. This problem is unique and unsolved in its ability to resolve references in the visual field *even if* it has no corresponding symbol for attachment. Saliency centered search usually leads to filter-based solutions that incorporate inhibiting activation to limit fixation at a particular location. No system as yet incorporates a form of multi-class recognition for the purpose of higher level indexing in its search repertoire. Just like a toddler has the ability to point and gesture at objects that he or she cannot name yet, the visual subsystem should be able to provide location indices despite a lack object class. The SHARE system implements a basic visual search feature that leverages a novel foreground reasoning mechanism to address foreground selection in situated computer vision (e.g., actively perceiving and making decisions about where to look next).
2. Intrinsic and innate biases: these play a critical role in reference resolution when the goal of the reference is unknown. A synthetic attention system that interacts with a biological attention system must address and capitalize on similar biases that show up. For example, red objects and motion play a role in visuo-foreground selection in humans. While many researchers are interested in identifying these biases, re-

searchers need more discoveries for a comprehensive and plausible system that integrates these biases. While some researchers believe these problems are unimportant to multi-robot joint attention (Kaplan & Hafner, 2006), in fact, these factors play a key role in systems that interact with biological agents (e.g., in human-robot interaction) that empirically capitalize on these biases. Key discoveries in bias-finding perform a function in integrated bias saliency systems and have led to biologically inspired, sensorimotor robots where saliency primarily directs the robot's gaze, and hence, its space of action. The system designed throughout this dissertation leverages the saliency models of others in the architecture.

3. A working memory system: this allows the visual attention system to dynamically tether unresolved lexical entities to its perceptual field while providing a memory system designed to store data on found objects from the visual field. The SHARE system utilizes a scoped memory system for just visual working memory.

Unique to this work and the research presented in the following chapters, the SHARE system unifies the previously mentioned factors into an agent architecture. Any visually guided situated system must address the problem of identification and selection. This chapter introduces a new metric that attempts to unify and make progress towards social-foreground selection. The metric also allows both recognition and saliency based approaches to be compared using a single function. This new metric forms the intellectual backing to discuss progress made in Chapter 4.

### **3.1 Relevant work**

Synthetic human-robot social learning systems must interact with their biological counterparts and while humans are significantly different, engineers can design robots to be interactionally compatible with them. How then does a researcher make progress in achieving shared attention? To resolve this challenge, it is instructive to consider the positive attributes of a synthetic shared attention system. Such a system allows the architecture to select relevant stimuli from its environment that its human counterpart directs it toward

(regardless of whether or not it understands the object it is guided toward) and it allows the underlying sensory stimuli to be used within a learning framework. To fully engage with the shared attention system thought experiment, researchers and engineers must also validate that the stimuli the human directs the robotic system to notice is, in fact, the correct set of stimuli or features. Section 3.2 discusses exactly this measurement and provides a new metric of success, allowing shared attention systems that rely on synthetic social-perception to trust their output.

Artificial intelligence systems still lag in representing high level phenomena; however, they are enjoying popularity in computer vision. These are the more recent foveated models in deep convolutional neural networks applied as complex classifiers (Stollenga, Masci, Gomez, & Schmidhuber, 2014; Xu et al., 2015) in cluttered scenes or the less popular attempts at incorporating elements of attentional leakage through a stacked array of global filters (Tsotsos, 2011). The human-robot interaction domain traditionally focuses on learning gaze following behavior itself rather than representing attention explicitly. Beginning with Scasselatti's work on gaze following (Scasselatti, 2001), much work has focused on learning the behavior itself rather than explicitly designing such behavior. Recent models integrate elements of active learning (Doniec, Sun, & Scasselatti, 2006) and reinforcement learning (Triesch, Teuscher, Deák, & Carlson, 2006). Doniec and Scasselatti's work is most relevant to the embodied nature of the problem. Their research shows clear evidence that the robot influences the participant and the robot follows the pointing gesture behaviorally. It is now evident that the robot has shared attention with a participant from a behavioral standpoint. However, has the system successfully chosen the correct stimuli in its perceptual field to utilize for the purpose of recognition and learning? In this researcher's opinion, the phenomena that are relevant are measurements of foreground and background (in the A.I. literature) or the figure-ground separation in neuropsychology literature. It appears there is no prior work in the context of human-robot interaction that asks the participants if the information they report is their figure-ground separation.

## 3.2 Shared attention : metric and hypothesis

Sharing attention with a human is equivalent to sharing a current world model at the current time step. In the situated approach, an agent chooses its frame by finding regularities in its environment that serve to advance the higher level goals the agent is attempting to achieve. This has been called the endogenous factor in the literature (Posner & Cohen, 1984). Vice versa, when higher level goals are not activated, the agent enters into a strict stimulus-response mode, called the exogenous factor. Artificial intelligence research that focuses solely on the exogenous factor typically falls under the topic of saliency while research that focuses on endogenous factors resides under the broad idea of visual routines (Ullman, 1996), visual search, guided search (Wolfe, 1994) or some other more ecological form of computer vision. The challenge now is to share this dynamic with another agent. Activating or stimulating objects and regularities in the shared environment to affect a participant's gaze and selection mechanism is difficult and the subject of many interaction studies interested in human-robot interaction.

This research focuses primarily on foreground reports as a different, quantifiable measure of sharing attention. By priming the participant's objective to share attention with a robot, *this work hypothesizes that a specific class of nonverbal gestures called deictic gestures will emerge and the goals of each participant and the scene being presented will shape the actions themselves and their parameters*. Regardless of the biological plausibility of the model the robot utilizes, the question is whether the human inhibits and activates similar perceptual foreground (based on self report) and whether the robot can adaptively shape its filtered foreground appropriately.

### 3.2.1 A new metric for sharing attention

Given any generative model that predicts the sensory foreground of the participant,  $F^p$  and given the reported foreground by the participant  $F^g$  from the same ego-centered camera system, the normalized mean squared error is probably the best way to understand whether or not attention is shared:

$$NMSE(F^p, F^g) = \frac{1}{wh} \sum |\hat{F}_{i,j}^p - F_{i,j}^g|^2$$

where  $i, j$  are pixel coordinates and  $w$  is the width of the image,  $h$  is the height. This assumes that the user selects the foreground from the robot's perspective after the fact.

If the robot is provided the goal to share attention with a specific foreground,  $F^{Goal}$ , then the user must provide an accurate prediction of that foreground, minimizing  $NMSE(F^{Goal}, F^g)$ . Vice versa, if the human is given a foreground to communicate, the robot must predict from its sensors the given foreground, minimizing  $MSE(F^{Goal}, F^p)$ .

### 3.2.2 Human dyad pilot study

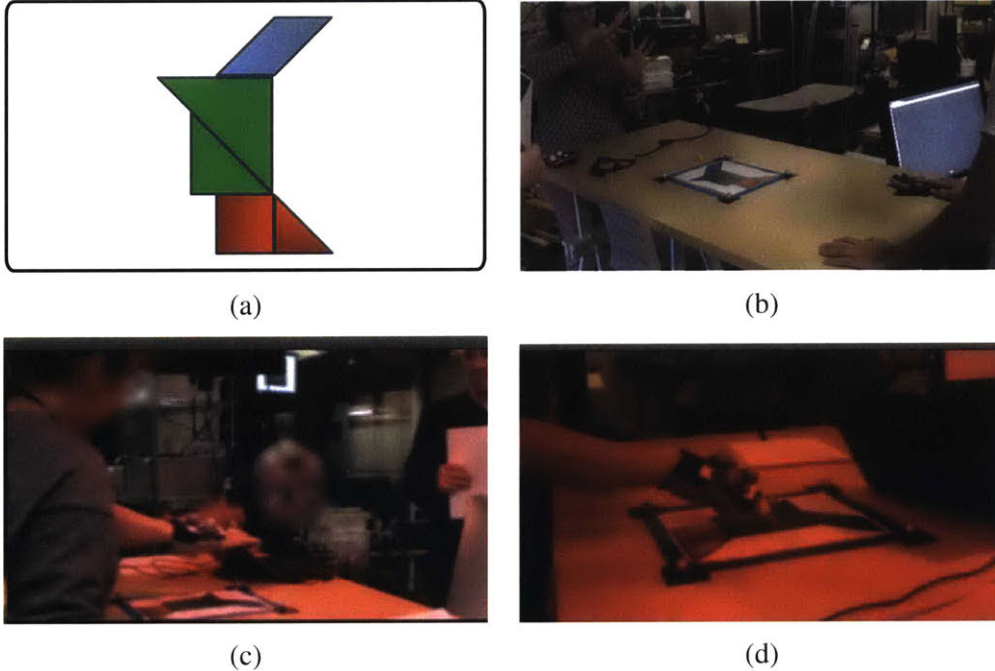


Figure 3-1: Observed deictic action and gesture during pilot. Tangrams offer a constrained domain for both selection and recognition. a) Demonstrates an example of a tangram. The observed gesture throughout the experiment was diverse. b) In this excerpt screenshot, iconic gesture (a rarely used referential act) was observed. Another gesture observed throughout the experiment can be found in c) where single extension, sweeping gesture were observed and d) where numerous, precise gestures were observed.

To observe this dynamic, 22 participants were recruited from the local MIT community

to communicate a foreground to the other participants. Ten dyads were formed to exchange gesture and report foreground. One early dyad was thrown out after results showed modifications of the instructions elicited iconic gesture (e.g., Figure 3-1). This class of gesture is characterized as mimicking the shape or dynamic of something in its environment (see Figure 3-1). Iconic gesture has been previously linked to higher level attentional effects such as priming that are not well represented in these perception models (Wu & Coulson, 2007).

One fundamental challenge in computer vision is to detect and use boundaries, the fundamental trade-off between aggregates of regularities making up an object vs. the regularities themselves. This work breaks down the exchange of nonverbal behavior toward sharing perception using the domain of tangrams. This domain allows one to arrange sub-component pieces to build higher level object structures and to represent parts of larger objects well. In addition, this artificial domain is easily represented on a computer screen to allow the participant observing the gesture to select the desired foreground on a touch screen.

Within the dyad, roles were given. One participant was told he or she would receive a goal to communicate to the other participant, specifically a foreground. The other participant was told he or she would observe and receive the foreground. The participants were told they had only 5 seconds to exchange gesture. This pressure forced the participants to exchange gestures quickly to resolve the ambiguity of a single gesture.

The communicator had a private screen that primed the participant with the task goal and the observer had a private screen to enter in the foreground. The real-life scene and the two computer screens were synchronized and the orientation was taken into consideration when presenting the screen. In other words, the participants stood side-by-side rather than opposite one another to share scene orientation (an important consideration for human perception).

### 3.3 Analysis of behavior

After being asked to communicate a specific foreground goal to the other participant without talking, one participant began to gesture using one of two major classes of intentional referential gestures: broad sweeping gestures and a highly precise pointing gesture. The participant used the pointing gestures to direct attention to the subcomponents of the foreground tangram, using many referencing actions to refer to the foreground. Other participants seemed to use broad sweeping gestures to indicate “all of the tangram pieces within this space” (see Figure 3-1). This may indicate the human’s ability to perceive the intended object that the tangrams compose. Similar to seeing the forest from the trees, or vice versa without representing trees (Greene & Oliva, 2009), *the participants referred to large aggregates as well as precise subcomponents.*

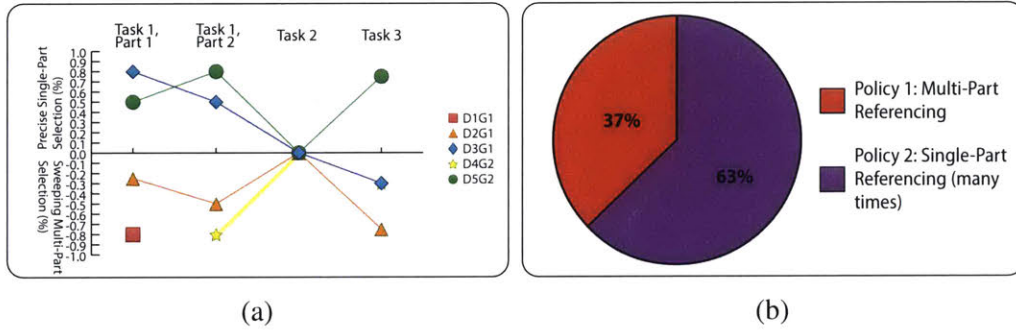


Figure 3-2: An analysis of deictic exchange: action class and numeracy. a) Dyad preference for reference strategy. X-axis: Scene (T1-4) Y-axis: +1: 100% general gesture, 0: equal gesture types utilized, -1: 100% precise gestures used. Each line represents a dyad (D1-5). b) Of those dyads that correctly communicated the foreground, this chart presents the fraction of participants who used the basic referencing policy.

The videos from the pilot study were video coded for gesture type and numbers. They are presented in Figure 3-2. It is clear that the variance the participants used ranges broadly and fairly evenly from the general gestures to the precise gestures. Task 3 was a special case scene in which ten tangram pieces were separated by a large margin of empty space between them. Participants reacted with only precise gestures so the task was thrown out for the comparison. There is corroborating evidence from (Saupé & Mutlu, 2014) that these general gestures show up frequently within referential behavior. When robots must execute deictic referencing behavior, Figure 3-2b shows that the robot may need to support



both multiple, precise referencing gestures as well as single, general sweeping gestures, depending on the underlying stimuli the robot is referencing. Generally, referencing behavior policies are adaptively switched within the interaction and perception systems must support many of these referencing behaviors. Additionally, the perception system that understands this referencing behavior must support classifying these foregrounds into recognizable objects. This important factor is explored more in Chapter 4. Once behavior is exchanged, the results show that the observer almost always is able to guess which foreground is being communicated (83%) by the participant's peer.

### **3.4 Discussion**

This study is designed to identify general patterns of deictic gesture exchange and map them to specific referenced regions of the visual field. While only a small number of participants were recruited, a brief glimpse of certain dynamics is evident. It is clear that given a communicatory goal for a human participant, peers can successfully exchange both deictic and iconic gestures (similar to those found in Figure 3-2) toward sharing foreground (Section 3.3). The work of vision scientists like Marr (Marr, 1982) and more recently Zhu et. al (Guo, Zhu, & Wu, 2007) postulates an internal mechanism that can sketch the shape of the image, so called the 2.5 sketch (for Marr) or the primal sketch (for Zhu). The work of primal sketch implementations is to represent not just the statistical patterns found in the image but also the shape, form and sketch of the scene or region itself. These underlying sketches may provide interesting models that become useful to the problem of resolving referential iconic gesture (e.g., that present in Figure 3-1, top right) through re-priming the vision system to find similar shapes in the environment. No system today can address the problems presented by this interaction (e.g., primal sketches, visual search and visual priming), let alone integrate them into a situated architecture. The following chapter attempts to build one critical component of this architecture: the ability to perform visual search within an image.



# References

- Burke, R., Isla, D., Downie, M., Ivanov, Y., & Blumberg, B. (2001). *Creature smarts: The art and architecture of a virtual brain*.
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.
- Doniec, M. W., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *6th IEEE-RAS International Conference on Humanoid Robots* (pp. 34–39).
- Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, 58(2), 137–176.
- Guo, C.-e., Zhu, S.-C., & Wu, Y. N. (2007). Primal sketch: Integrating structure and texture. *Computer Vision and Image Understanding*, 106(1), 5–19.
- Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2), 135–169.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.
- Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. *Attention and performance X: Control of language processes*, 32, 531–556.
- Sauppé, A., & Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 342–349).
- Scassellati, B. (2001). *Foundations for a theory of mind for a humanoid robot* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

- Stollenga, M. F., Masci, J., Gomez, F., & Schmidhuber, J. (2014). Deep networks with internal selective attention through feedback connections. In *Advances in Neural Information Processing Systems* (pp. 3545–3553).
- Tomasello, M. (2000). *The cultural origins of human cognition*. Harvard University Press.
- Trevarthen, C. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. *Before speech: The beginning of interpersonal communication*, 321–347.
- Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, 9(2), 125–147.
- Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press.
- Ullman, S. (1996). *High-level vision: Object recognition and visual cognition* (Vol. 2). MIT Press. Cambridge, MA.
- Wolfe, J. M. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2), 202–238.
- Wu, Y. C., & Coulson, S. (2007). Iconic gestures prime related concepts: An ERP study. *Psychonomic Bulletin & Review*, 14(1), 57–63.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., . . . Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3), 5.

## Chapter 4

# Design of a directable artificial visual search architecture

Visual search is described by Jeremy Wolfe (Wolfe, 1994) as eye behavior that utilizes a limited, fixed capacity towards identifying a set of objects in a serial manner under a retinopic fixation point. This visual search behavior is not well understood but there are actually many models of visual search that are plausible and well studied in psychology literature (Wolfe, 1994; A. M. Treisman & Gelade, 1980). This chapter provides a model of synthetic visual search that indexes its world based on a flexible aperture that recognizes objects through perceptual grouping of the considered foreground. Vision scientists have yet to solve the problem of building a visual search architecture to support this ongoing behavior in a way that competes with standard computer vision metrics. But progress is being achieved in limited capacities as we'll see in Chapter 5 which explores an extension of past models that are competitive (Mnih, Heess, & Graves, 2014) with standard convolutional neural networks (LeCun et al., 1989) on the same dataset. Regardless, the sets of basic mechanisms required to implement a system that can perform serial and contextual recognition go beyond the current state-of-the-art in computer vision. The SHARE architecture extends the abilities of previous models to allow an agent to share gaze and recognition synthetically with a human partner.

A situated gaze following mechanism can be implemented in a number of ways. One solution is to segment the scene into a number of regions followed by recognition (see

(Girshick, Donahue, Darrell, & Malik, 2014)) of the regions. This approach is prohibitively expensive due to its processing extraneous parts of the scene not pinpointed for recognition. This pruning can be performed at an earlier stage of the process with the region providing the label and index that the human partner has selected. Alternately, one can take a different and more computationally efficient approach: to build a perception system that can reason directly about foreground (i.e., region) recognition and provide the ability to follow both gaze and gesture toward an indexed target while simultaneously inhibiting the process of recognizing the rest of the scene. There is a huge challenge in supporting the types of reference goals that one may refer. This chapter (specifically Section 4-2) provides evidence that semantic segmentation may not provide a sufficient substrate on which to resolve exophoric references.

The goal of this work is to build a flexible perception system that can be used in human-robot interaction domains to extract and learn about select environments through human interaction. Design of such a system depends on requirements relevant to joint attention (Tomasello, 2000) in which a robot may be directed to attend to something new and still have the capability to refer to and learn from this sensory experience. While the SHARE system also generates goal-oriented deictic action (critical to directing a partner's gaze), this chapter explores the performance of pixel-level referencing vs. indexical (i.e., object-oriented) referencing. In essence, the approach is to extract deictic indices through the integration of information across multiple modalities via interactions with both the environment and a social partner. This triadic relationship between social partner, self and environment sets the stage for a more situated approach than traditional robot-environment interactions alone.

#### **4.0.1 Simulating a peer's reference to improve exophoric resolution**

One unique aspect of this work that will be further explored in Section 4.3 is the underlying model that puts internal modules in competition to select and estimate foreground. Models of competition show up in a number of cognitive architectures. Fieldman and Ballard described these early on and hypothesized that connectionist models were impor-

tant (Feldman & Ballard, 1982). Scientists later applied competition to selective attention, (Koch & Ullman, 1987) popularized in robotics through the models of behavior-based architectures (R. Brooks, 1986; Blumberg, 1996; Burke, Isla, Downie, Ivanov, & Blumberg, 2001). Competition offers a controlled way to mediate coordinated parallel architectures while taking advantage of the best of the trade-offs of all the submodules. Rather than focusing on popular models, competitive learning in artificial neural learning or recurrent networks, Section 4.3.1 proposes a new method to resolve competition and tests it against the dataset.

Competition must be resolved in a meaningful way in an internal visual system where multiple systems contributing to the foreground generate a number of proposals. Given an image and a reference, which underlying system best estimates the referenced foreground? The solution is to build an internal model of perceiving and referencing, both as a spotlight model, and align perception and reference to the same stimulus. A visual search mechanism that points and fixates on a particular location to identify its target seems like a natural match to handle deictic referencing. The foreground is back-projected from the indices and can produce a unified way of generating pointing positions, both proximal and distal, using simple geometry. These referencing actions represent a self-as-simulator model that allows the system to compare incoming referential actions with a most-likely foreground reference and potential index. This novel mechanism is described further in Section 4.3.2 and is a part of the hypotheses described in the following section.

## **4.0.2 Conditions and hypotheses**

Based on previous findings from Chapter 3, this research hypothesizes that the precision of foreground being shared between a human and robot is highly dependent on the underlying model of attention. Namely, the following hypotheses are stated:

- H1: When referencing objects that are already known to the synthetic agent, a goal-directed visual search mechanism with the capability to recognize objects and produce boundaries estimates a more precise foreground than simple saliency models.
- H2: When referencing objects are unknown to the synthetic agent, a model of rec-

ognized objects being back-projected onto the image is insufficient to model the referenced foreground compared to saliency models that could estimate the foreground best.

- H3: Putting these models in competition to take the best of both models allows the system to perform exophoric reference resolution and produces ideal results for all potential goal foregrounds.

The success of H3 is important for opportunistic and embodied learning. If a robot is designed with a signal that an exophoric reference maps to a location with an unrecognized foreground template, then a potential opportunity to label the visual template would emerge from the interaction. In addition, Chapter 2 explores examples of how these templates can be sourced from on-line participants using a frame alone.

## **4.1 Related work**

For a robot to share attention with a human participant and vice versa, it needs to take actions in the world to affect its partner’s visual system. Evidence from (Doniec, Sun, & Scassellati, 2006; Bainbridge, Hart, Kim, & Scassellati, 2008) suggests that robotic nonverbal deictic gesturing affects the gaze trajectories of human partners. Yu and Smith also present evidence that joint attention occurs through multiple sensory-motor pathways, engaging with both referential and gestural behavior as well as gaze trajectory manipulation (Yu & Smith, 2016). For these reasons, the system presented herein internally represents a set of deictic *pointers* (e.g., an origin-ray-scope model) that helps the system keep track of location indices in a local memory. This section will cover related work in attention, robotics and deictics.

### **4.1.1 Joint attention and attention in robotics**

The work of Scassellati on gaze following and sociability led both to become core issues for social robotics (Scassellati, 2001). His work focused on a number of mechanisms that would be required to implement a deeper theory-of-mind for humanoid robots, inspired



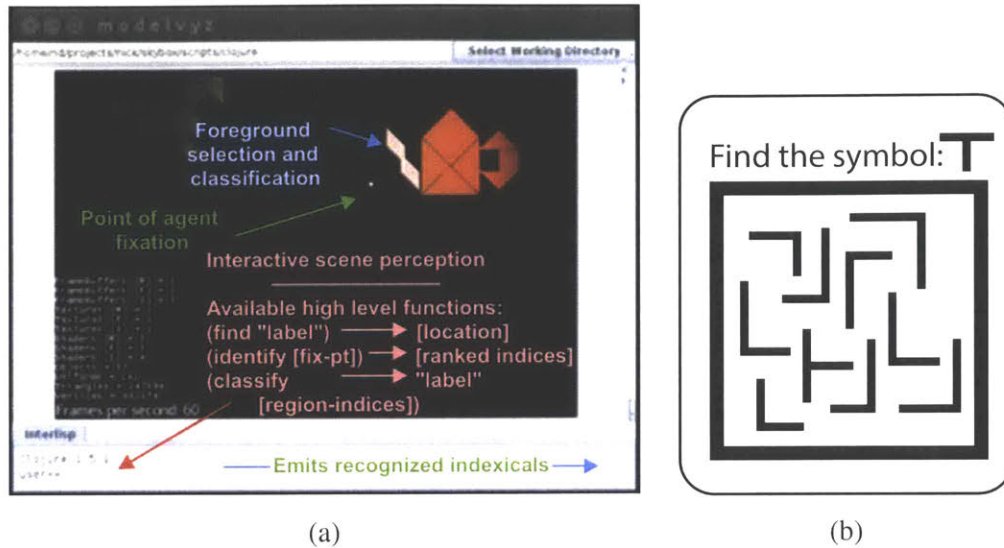


Figure 4-1: An illustration of synthetic and natural visual search. a) The SHARE system: tangrams in view with foreground being selected and identified as a known “whole object.” The dataset was collected from on-line contributions. Functions available are “find,” “identify” and “classify.” b) A demonstration of the reader’s human visual search adapted from Wolfe (1994). The objective is to find the “T” symbol within the image. Notice how your eye searches the image in no particular order.

heavily by the work of Butterworth (Butterworth & Cochran, 1980) and Baron-Cohen (Baron-Cohen, 1997). The joint attention system in this cohesive system followed gaze and pointing gestures toward a target location but was unable to recognize the object under its fixation point. Following this work, effort began on learning to follow gaze from a developmental perspective. Nagai et. al., Doniec et. al. and Triesch et. al. (Nagai, Hosoda, Morita, & Asada, 2003; Triesch, Teuscher, Deák, & Carlson, 2006; Doniec et al., 2006) are all concerned with how to map referential gesture or gaze to predetermined objects in the world, or in other words, *learning to follow* a referential gesture toward objects with an existing model. The questions investigated in this section expand on previous work by incorporating elements of recognition at fixation locations as well as requesting self-report foregrounds from human participants. One evaluates the success of this system by measuring the error of the reported *foreground* (a mapping of inhibited factors and uninhibited factors) and the predicted foreground of the shared factors.

As discussed in Chapter 1, visual attention systems can be roughly characterized as bottom-up or top-down. Bottom-up approaches focus primarily on filter stacks (Tsotsos,

2011; Itti, Koch, & Niebur, 1998) due to their ability to facilitate leakage while top-down attention systems emphasize recognition systems that drive movement of the fixation to the next point. Driving fixation behavior using conspicuity maps is typically carried out by inhibiting observed locations and reweighting the same map to reveal new locations to target. Recognition systems focus primarily on localizing objects within images through some search mechanism (see Figure 4-1). Both approaches clearly affect gaze behavior and both are incorporated in the SHARE system section presented later in this chapter.

Since both bottom-up and top-down approaches help characterize the gazing behavior, this chapter will touch on both factors and discuss relevant work in this area. One system that unifies saliency maps as a means to extract templates and learn new models on a roaming robot is Frintrop's VOCUS attention system (Frintrop, 2006). VOCUS learned about objects via saliency maps and a curiosity system driven to find new and novel objects in its world. VOCUS also focused primarily on object-environment relations and was not biased to learn from other agents in its environment. A coverage of computational attention would not be complete without the decades-long research of Tsotsos (Tsotsos, 2011). He presents one theory of computational visual attention based primarily on Gelade and Treisman's model of visual search (A. M. Treisman & Gelade, 1980) that can compute saliency values for arbitrary images. This work emphasizes the computational mechanisms surrounding attention itself and does not account for social factors. None of these algorithms focus on robotic joint attention in which the robot plays an active role in sharing attention with a human participant.

The process of shared attention in this chapter examines the exchange of social cues to generatively and collaboratively design the shared foreground map (DePalma & Breazeal, 2015b) rather than the learned development of gaze-following toward behavioral joint attention.

### **4.1.2 Shared attention through a common representation: deictics**

Referential gesture that is dependent on context is sometimes referred to as deictic. Referential gesture and deictic use in human-robot interaction has been studied in various ca-

capacities. (A. G. Brooks & Breazeal, 2006) presents a model of multi-modal deictic generation in communication that leverages grammar models to generate deictic gesture. Saup   (Saup   & Mutlu, 2014) presents work on specifying a number of categories of deictic use in reference which include pointing, presenting, touching, exhibiting, grouping and sweeping. Other effects such as synchrony (Rolf, Hanheide, & Rohl  ng, 2009) and motionese (Nagai & Rohl  ng, 2009) may also contribute to the intentional capitalization of innate biases that drive saccade behavior. Additionally, appropriate deictic reference use has been attributed to improved collaboration between humans and robots in difficult tasks (Admoni, Weng, Hayes, & Scassellati, 2016). Though joint attention has been studied in various capacities under different definitions, (Kaplan & Hafner, 2006) convincingly argues that true joint attention is an intentional, goal-oriented process and models that leverage saliency systems where innate biases serendipitously grab the attention of the interaction group should be considered unintentional shared attention. This chapter focuses primarily on foreground-as-goal and utilizes deictics as a communicatory action to synchronize foreground.

For a robot to share attention with a human participant and, vice versa, it will need to take actions in the world to affect its partner’s visual system (Bainbridge et al., 2008; Yu & Smith, 2016). Additionally, the robot will need a sensory system that can handle the actions that are directed at the robot. This researcher drew inspiration from a number of sources when researching this seemingly simple question, from psychology and human-robot interaction to state-of-the-art robotic attention systems. This chapter documents progress made toward a visuo-motor shared attention system that simulates others to improve foreground sharing.

## **4.2 Problem: the socially driven visual search task**

Many tasks in static environments require saccading from position to position while recognizing and indexing the environment being observed. As described earlier, this process of visual search may be significantly more social than previously thought. Research in joint attention argues that these processes may be deeply social in nature. In this work, joint or



Figure 4-2: Some observed behaviors during a pilot experiment when attempting to direct the attention of another human participant. Left: a participant used bounded hand gestures to refer to the space between the palms (highlighted in blue). Right: precise pointing meant to highlight one particular region that must be interpreted to be one piece of the tangram figure (highlighted in blue).

socially guided visual search is defined as the collaborative process in which synthetic embodied agents exchange gesture toward sharing foreground (e.g., parts of the scene that are activated and not inhibited). The foreground is a measurement space which may highlight object silhouettes or the underlying pixel saliency itself (see Figure 4-6). An integrated approach to improving socially driven visual search for agents will require internal robotic processes to handle sequential classification (or recognition) of scene indices in the environment. In the case of socially driven visual search, the agent will index the scenes together with a social partner.

#### 4.2.1 Task requirements

Collaborative human-robot joint attention requires the visual search task to incorporate both the ability to predict the participant's foreground and take goal-oriented action toward changing the interaction partner's direction of gaze. Foreground is defined as binary maps that represent the thresholded salience of the scene. This prediction allows the robot to define a shared attentional space on which both the human partner and the robot may learn. Expanding knowledge through attention mechanisms is one of the key ways that learning progresses in biological agents. The investigation of key social action tools like gesture found to affect gaze of human social partners has been conspicuously missing from robotics literature until recently. Any system that intends to affect the gaze of a partner will need to take real world deictic action to direct its partner's gaze. Deictic action has been shown

to affect gaze meaningfully in work by other researchers (Bainbridge et al., 2008; Sauppé & Mutlu, 2014). Evidence successfully observed in dynamic scenes is presented later in Chapter 6.

A system that can support the demands of socially driven visual search will require advances in computer vision and interaction design. One issue impeding progress is the lack of a key metric of success that will allow researchers to compare and incrementally improve progress in human-robot joint attention systems. Chapter 3 attempted to rigorously argue that the socially driven process should be compared across systems to improve one key metric that would enhance dialogue, collaboration and learning systems: namely foreground. *Sharing foreground* is in effect sharing an *observation space*. To measure success of sharing foreground, a normalized mean squared error metric is proposed by DePalma et. al. (DePalma & Breazeal, 2015b) and described in Chapter 3. This metric allows shared attention systems to compete on equal footing and compare systems that may have radically different models relying on either recognition-based indexing or simple saliency maps. Additionally, this metric may enable perceptual learning through simple interaction alone. While systems like VOCUS (Frintrop, 2006) involve learning to index from simple saliency maps, social learning could improve these systems by capitalizing on social interactions and allowing these systems to compare visual-learning models across domains.

To measure the success of a solution to this problem, the task should have the following qualities:

- It should have the complexity to compose and aggregate sub-figures and wholes so visual recognition and indexing become core phenomena for investigation.
- It must have the capability to reference non-object type foregrounds (Figure 4-6 reveals one mystery).
- It must also allow the human participant to report his or her goal foreground to the robot.

Since tangrams offer the necessary features, they were used for the purposes of these experiments. Simple tangrams ensure clean templates and boundaries that can be employed within the contexts of both learning and referencing.



### 4.3 Computational model

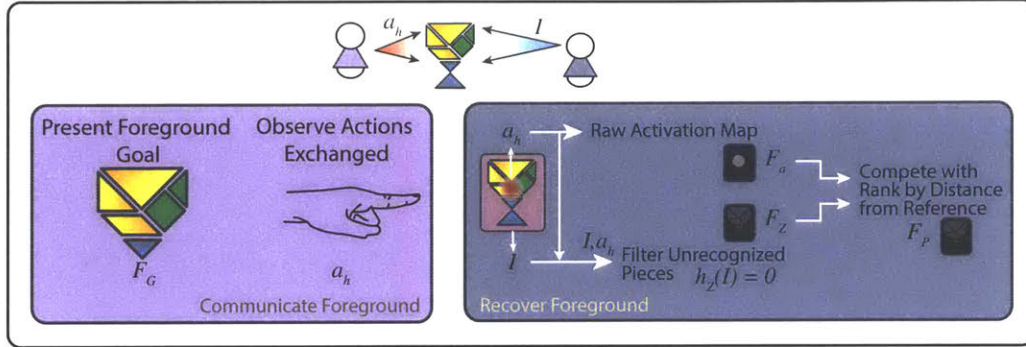


Figure 4-3: An illustration of capturing the deictic reference to use for selection and hypothesis enumeration.

Investigating how indexing and recognition, saliency and peer simulation affect foreground sharing required designing a three-factor foreground contribution system, described in detail in Sections 4.3.2.1, 4.3.2.2 and 4.3.4 respectively. A sketch of the problem overview in Figure 4-3 explains the system and roles of the exchange. In this overview, the role of the director is to communicate a goal foreground to the viewer. A goal foreground  $F_G$  requires an action  $a_h$  (in this case, defined as coming from a human,  $h$ ). The objective of the viewer is to infer which foreground is being communicated. This model requires the following input: the image of the scene,  $I$ , and the action in continuous space  $a_h$ . The viewer must make sense of the reference and its mapping on the environment, extracting just the foreground that is relevant. This process has been called exophoric reference resolution in discourse analysis ((G. Brown & Yule, 1983), Chapter 2).

Using a shared scene made up of tangrams, the system attempts (at a pixel level) to predict the goal foreground provided to a user of the system. Roles are assigned to each participant in the dyad: the observer of the action must predict the goal foreground of the director and the director must provide a set of actions that the observer sees. Since the study is primarily focused on the observer role for this chapter, the only information provided to the robot is the image of the scene and the referential centers of the deictic actions. In Figure 4-3, a goal foreground,  $F_G$  is given to the director (i.e., the chalice top in the figure) who then provides a set of actions,  $a_h$ , for the observer. With only the actions from

the human partner,  $a_h$  along with the image  $I$ , the robot observer and the algorithm must output a binary foreground prediction  $F_P$  from each of the underlying subsystems described in the following sections. The indexing subsystem recognizes subfigures and wholes, then chooses the most likely referred candidate,  $\langle F_Z \rangle$  where  $Z$  represents the fact that it is the result of a set of binary classifier functions  $h_Z(I)$  operating on different subsections of the image. Additionally, deictic action is mapped directly to the scene to produce a secondary, binary pixel map to produce  $F_a$  that competes with the index to support referencing of both known and unknown objects. Through a process described better in Section 4.3.4, these hypotheses compete to predict the foreground that the human participant is attempting to direct the robot to observe.

To better understand the performance of this system, one compares a number of approaches to evaluate their effects on the visual attention system. The main goal was to shape the foreground between the human and the robot with a specific interest in the deictic class of signaling. Within this class, a referential gesture must be detected and spatially resolved. The computational model consists of three components: 1) a visual search mechanism that uses previously learned indexing and recognition to resolve the reference (Section 4.3.2.1), 2) a simple flashlight model interpretation of the deictic action to highlight the appearance of the image itself, given a cone model of referencing (Section 4.3.2.2) and finally 3) a system that allows the previous subsystems to compete through a thresholded confidence value generated by a novel perspective-taking mechanism (Section 4.3.4). The following subsections will cover the details of each method, respectively.

### 4.3.1 Competition algorithm

Using a novel self-as-simulator approach, this system resolves competition that occurs during the unification of two different foreground prediction subsystems. These subsystems are called *proposal pumps* or just *proposers* meant to describe the asynchronous nature of these subsystems and the snapshots that emerge during incoming deictic action events. The system attempts to hypothesize potential foregrounds used as input to generate potential deictic actions the robot can take,  $a_r$ , which are used to compare against the incoming ob-

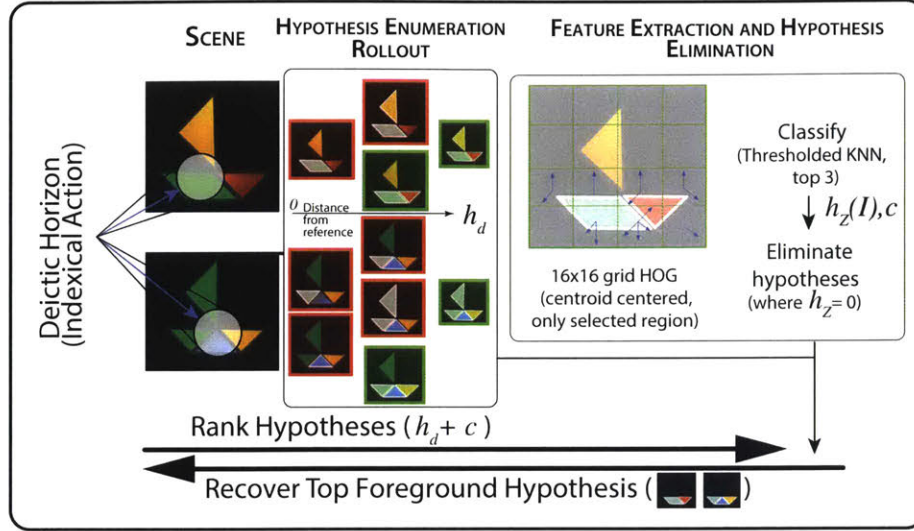


Figure 4-4: Illustrating hypothesis roll-out of known and recognizable wholes from a reference. Horizon of roll-out is given by how proximal or distal the reference is to the location.

served actions  $a_h$ . The internally simulated actions that the robot takes to communicate a foreground helps the robot predict the most likely candidate foreground being conveyed. Recognized indices produce high confidence values that the reference index is known to the robot and is a simple reference, while new foreground stimuli from the saliency system are routed to a learning mechanism to perform further learning.

### 4.3.2 Interpreting object-directed, referential action

Interpreting the object that the robot should attend to involves two complex processes: 1) identifying possible objects that are well known or recognized and 2) identifying regions that the human intends the robot to attend to that have no recognizable attached index. The following subsections detail how these two competitive forces compare on foreground activation within the model.

#### 4.3.2.1 Proposer 1: Reference-to-object effects

This particular section is documented in published work (DePalma & Breazeal, 2016b) which discusses this particular subsystem's contribution to the robot's visual attention system. The pipeline is shown in Figure 4-4. First, referential action is specified as having an



aperture-point estimate  $a_h(\bar{x}, \bar{r}, \theta)$  in space  $\bar{x}$ , a vector direction  $\bar{r}$  and a range (angle  $\theta$ ) of affected foreground. With known objects  $Z = \langle z_1, z_2, \dots, z_n \rangle$ , the indexing recognition system can classify a current foreground hypothesis as a known part or object. Multiple object classification is performed by first selecting a number of foregrounds and then attempting to classify these foregrounds using a multi-class classification technique.

The algorithm begins by using an aperture-point estimate  $a_h(\bar{x}, \bar{r}, \theta)$  to project a conic horizon in which to enumerate the object hypotheses during the search for known objects (see Figure 4-4). A number of foregrounds are selected by enumerating all combinations of tangram pieces whose center points  $c_Z$  are within the projected ellipse from  $a_h(\bar{x}, \bar{r}, \theta) \rightarrow c_a$ . For each foreground hypothesis, the foregrounds are filtered where the trained binary classifiers are restricted to  $h_Z(F) = 1$  for the label  $Z$ . When all of the possible labels are classified in the given reference region, ranking occurs using simple inverse distance  $d_i(c_Z, c_a) = \frac{1}{c + \ell_2(c_Z, c_a)}$  where  $\ell_2$  represents the  $\mathcal{L}^2$ -norm.  $c_Z$  is calculated by taking the centroid of the foreground in which  $h_Z(F) = 1$ . HOG features (Dalal & Triggs, 2005) from the rastered images of the tangrams are used to train the  $h_Z$  classifiers.

HOG features are computed by first convolving a set of orientation kernels to produce eight gradient images and then binning the results. Orientation intensity is computed using the convolution of 8 different kernels representing the 8-point cardinal and inter-cardinal directions. For each direction of the 8 orientation images,  $O_d$ , the magnitude represents the sum of intensities within that bin  $f_b = \frac{1}{wh} \sum_{i,j}^{p_1 < i, j < p_2} O_{i,j}$  where  $w$  is the width of the bin,  $h$  is the height and the bounding box is bounded by  $p_1$  at the top left, and  $p_2$  at the bottom right. For this experiment, the foreground is scaled to a 320x320 image and binned into 16 regions representing the classification feature space of  $f \in R^{16}$ . For each exemplar and label, this process reduces the image to this 16 dimensional feature space and performs classification using KNN classifiers to select the most likely label.

The object labels that emerge from the model  $Z$ , are then reprojected back and re-rastered to produce a foreground image that is grayscale and thresholded to produce just the foreground that underlies the symbol. The output of this system makes up the output set  $\langle F_Z \rangle$ . The top ranking prediction (with smallest error),  $\hat{F}_Z$ , is used as the indexed and recognized contribution.

#### 4.3.2.2 Proposer 2: Reference-to-foreground effects

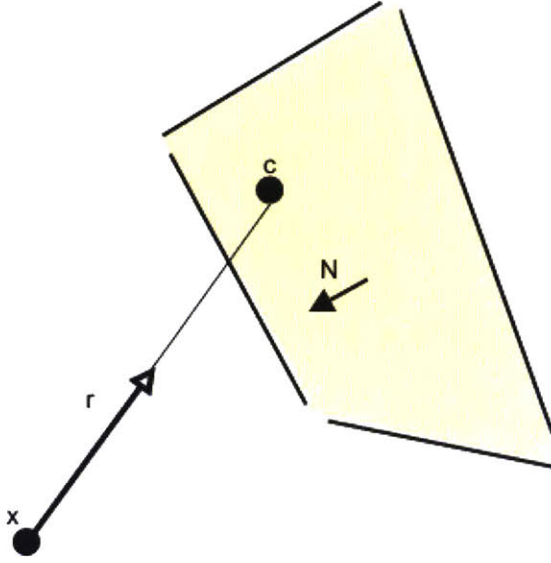


Figure 4-5: Computing the reference point against a surface from a distal location.

The point  $c$  is computed in world coordinates using the standard ray-plane intersection formula:  $c = x + rt$  where  $t = -\frac{(x \cdot N)}{r \cdot N}$ . The coordinates are normalized and scaled into image coordinates. Quaternion rotations of the direction,  $r$ , help raster the ellipse onto the image before a flood-fill algorithm completes the foreground selection. Once the foreground is computed, a nominal action estimation is used for further comparison by taking the computed foreground and using the method described in the following section.

In the event the system can't recognize the object at the point of fixation, the system should rely on a separate proposal. This proposer must only work on feed-forward functions across the entire image and may offer a chance for the rest of the architecture to learn a label from either people nearby or on-line (Chapter 2). This proposer offers just a simple foreground to label from the real world by projecting an ellipse of the action  $a_h(\bar{x}, \bar{r}, \theta)$  onto the image,  $I$ . Figure 4-5 describes the problem.

#### 4.3.3 Generating referential action

Work presented in Chapter 3 included observations of two situations that the system will need to handle (DePalma & Breazeal, 2016b, 2015a): precise pointing gestures and sweeping open-palm gestures. Sauppé and Mutlu (Sauppé & Mutlu, 2014) provide supporting evidence of the phenomena as well. Simulating referential and spatial deictic action requires an effort to generate similar referential gesture. To generate this action, the referential action generation system takes a simple foreground and generates a cone that covers the foreground on the image  $a_h(\bar{x}, \bar{r}, \theta)$ . Since  $\theta$  is global, the position and ray must be

calculated from the foreground.

To compute the action position  $x$ , and ray position  $r$ , a simple geometric model is used. The common point-ray model of reference uses the centroid  $c$  of the bounding box along with the width and height of the box to determine the radius of the point on the object's surface. The distance from the gesture is calculated to be proportional to the inner circumscribed circle of radius  $r$  of the foreground blob. The distance  $d$  is specifically calculated as  $d = \frac{r}{\tan^{-1}\theta}$  where  $\theta$  is the angle of influence of the referential gesture. Finally, the point and ray is  $\vec{x} = [c; d]$  with  $\vec{r} = [0; 0; -1]$  for each blob. Regardless of whether the agent wants to view the pieces of the tangram as a whole or point to each piece as a separate foreground, the agent can still reference either foreground map with the same simple algorithm.

#### 4.3.4 Fan-in unification

The model of competition used to resolve the two proposal pumps is to compare the incoming action model  $a_h$  with the generated deictic action  $a_r$  and to use the distance as a metric of comparison. The system compares the differences by examining and thresholding the expected utility of a contribution from one of the pumps that would be generated if the robot were to take the same actions in the participant's place with a particular goal foreground hypothesis. First, the hypotheses are enumerated (Figure 4-4) and a set of actions is generated,  $a_r$  for each hypothesis  $\langle F_Z \rangle$ . Using those actions, the two action policies are compared and scored by first computing the pairwise distances between the two point sets:  $\langle d(a_h, a_r) \rangle = \frac{1}{wh}(\|x_{a_h} - x_{a_r}\|)$  where  $\|\cdot\|$  is the  $\mathcal{L}_2$ -norm. Likewise, the system performs the same operation for the saliency contribution foreground:  $F_a$ . Finally, to bin the contribution of the indexical system, the findings show the expectation of the median values is the most predictive value of the indexical contribution:  $E[D] = E[\langle d(a_h, a_r) \rangle]$  where  $\langle d(a_h, a_r) \rangle$  is the median of the distances of each foreground set  $\langle F_Z \rangle$ . Finally the expectation is thresholded empirically by  $E[D] < 0.3$ .

## 4.4 Data collection

As reported in (DePalma & Breazeal, 2016b), tangrams were collected on-line over the period of a year. These tangrams were used to train the algorithms indexical references. Following this data collection, a number of participants were brought in to interact with the robot, Maddox. Maddox was a 56 degree of freedom humanoid robot with two arms for gesturing and a head with actuated eyes for gazing. Figure 4-6a shows the crowdsourcing interface used to collect tangram figures to be shared between the participant and the robot. The dataset included a total of 43 tangram scenes collected from the Internet, broken down into 125 subfigures to make a total of 168 possible reference labels to figures. In addition, the interface afforded a secondary task to random users on the Internet in which they were asked to highlight either the low-level regions of the figure that they found most interesting or to select the parts of the figure that can be considered 'separable.' For instance, on-line contributors highlighted legs, wings or hulls of boats and labeled them appropriately.

In addition, Figures 4-6e and 4-6f show the interface used to share the tangrams between the human and the robot. This interface allowed the robot to take actions toward sharing attention and in turn, the human was able to gesture back. The actions were captured on a device similar to Figure 4-6c. A total of 15 interactions were collected across 10 scenes, resulting in a total of 150 total scenes in which interaction was observed. Goals were provided to the participant on a separate screen (see Figure 4-6, middle row). The actions collected make up the entirety of  $a_h$  which were used to generate the predicted foregrounds from the visual attention system.

Simple tangram figures were given to the participant through a simple Pepper's Ghost illusion and instruction was provided. A scene that was collected on-line was presented to the dyad and roles were also given to each participant. The setup included a shared scene composed of tangrams. The participant was told that he or she had the role of *showing* the robot which part of the object the robot must select and then the robot would print out an answer on the screen (hidden from the user). No verbal communication would be exchanged.

The Leap Motion collected the gestures and began to stream the estimated positions


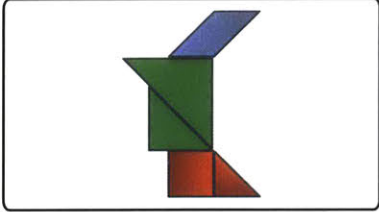


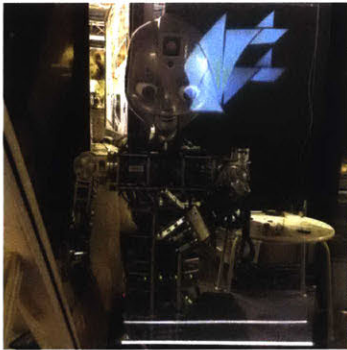

<p>Data Collection and Tangram Objects and Labels</p>	<p>BUILD TANGRAM OBJECTS</p> <p>This project aims to track the robot about novel objects that can be built from tangrams. To participate in this, please click the link below and follow the instructions.</p> <p>Instructions:</p> <ol style="list-style-type: none"> <li>1. Click the link below</li> <li>2. Select a shape at the top and a color below</li> <li>3. Click "Create shape"</li> <li>4. Hold Shift and click to rotate. Moving the mouse right/left rotates the shape clockwise/counterclockwise</li> <li>5. Click and drag to move the shape without rotating</li> <li>6. Once you are finished:             <ol style="list-style-type: none"> <li>1. Name your object in the text box below</li> <li>2. Click on "Save state"</li> </ol> </li> </ol>  <p>(a)</p>	 <p>(b)</p>
<p>Goal Foregrounds Collected</p>	 <p>(c)</p>	 <p>(d)</p>
<p>Data Collection Interface</p>	 <p>(e)</p>	 <p>(f)</p>

Figure 4-6: On-line and real world data collection used for collecting recognizable wholes and in person gesturing. Top row: a) Tangram collection web interface. b) Tangram figure taken from the dataset collected on-line. Middle row (both captured from the dataset): c) Object-based foreground goal collected on-line. d) Pixel-based goal foreground collected on-line. Bottom row: e) Pepper's ghost illusion used to situate the tangrams between the participant and the robot. f) Lower portion of the illusion device that uses a Leap Motion mounted on left side to collect gestures toward the illusion.

and orientations when the index finger was extended. The position and ray were collected from the proximal phalange (the “knuckle”) estimate and the ray was projected extending from the proximal phalange to the tip of the index finger (the distal phalange) without going through the intermediate bones. The origin of the bottom left of the screen (the tangram image was full screen on the monitor) and the normal vector were calibrated in Leap Motion space prior to interaction.

## 4.5 Results

First, the goals were separated into two datasets: low-level, unknown goals (UG dataset) and high-level, known goals (KG dataset). Next the set of actions was collected from the human dyad dataset and analyzed. Findings show that within a single dyad session, either one single gesture was exchanged (SG) or multiple referential gestures were exchanged (MG). For each of those groups, the datasets were divided into two groups: those in which the system had an object that it could predict and those in which the foreground was pixel-based. Dividing the databases allowed for comparison of the advantages and disadvantages of each situation that a robot might encounter. Error between the predicted foreground,  $F^P$  and the provided goal foreground,  $F^G$  using the normalized mean squared error (NMSE) is reported below. Again, Chapter 3 provides a full discussion of why this was chosen.

$$NMSE = \frac{1}{wh} \sqrt{\sum_{i,j}^{i,j=1..w,h} (F_{i,j}^G - F_{i,j}^P)^2}$$

As expected, H1 and H2 predict in Section 4.0.2 that isolating just the indexical system cannot provide reference resolution for all referential actions toward arbitrary goals. Figure 4-7 and Table 4.1 report the results of the robot’s ability to predict the foreground that the human participant is attempting to communicate. The goal image foreground and the predicted image foreground are aligned and the normalized mean squared error is reported on the y-axis. Figure 4-7 shows a clear performance difference between single gesture foreground prediction and multiple gesture foreground prediction. Reported  $p$ -values using a



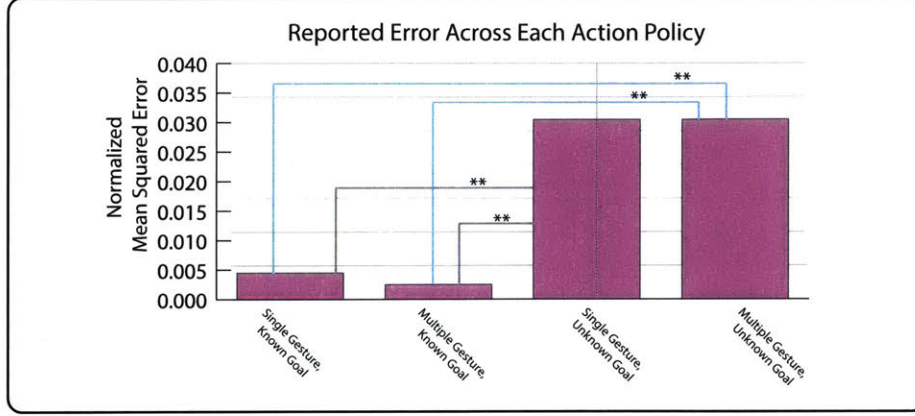


Figure 4-7: Foreground prediction performance of the visual search mechanism compared against multiple observed action strategies by the participant. The results show that the designed visual attention system performs well while resolving goal foregrounds which the system has already encountered and can recognize (known goal). However, it does not predict the foreground goal well for objects it has never encountered (unknown goal) for both action strategies. Reported significance in this figure are  $p < 0.01$  but refer to Table 4.1 for more detail.

	Reported Significance			Average NMSE
	MGKG	SGUG	MGUG	
SGKG	$p < 0.06$	$p < 0.001$	$p < 0.001$	0.004
MGKG		$p < 0.001$	$p < 0.001$	0.001
SGUG			$p > 0.9$	0.03
MGUG				0.03

Table 4.1: Reported significance and error of a single foreground proposer method against known goals and unknown goals. Significance values are reported using a Student's  $t$ -test, and the average normalized mean squared error is reported for each dataset on the far right.

Student's  $t$ -test between each group show that the visual proposal pump is not enough to predict foreground for all goals. In addition, an agent needs to balance foreground prediction between well-known object predictors and other approaches to handle pixel referencing. One example of note is Figure 4-6d, in which on-line participants wanted to reference edges of the image itself. Without explicitly modeling the edges of an object, it becomes a challenge to refer to such an entity with standard approaches.

#### 4.5.1 Analysis with competition

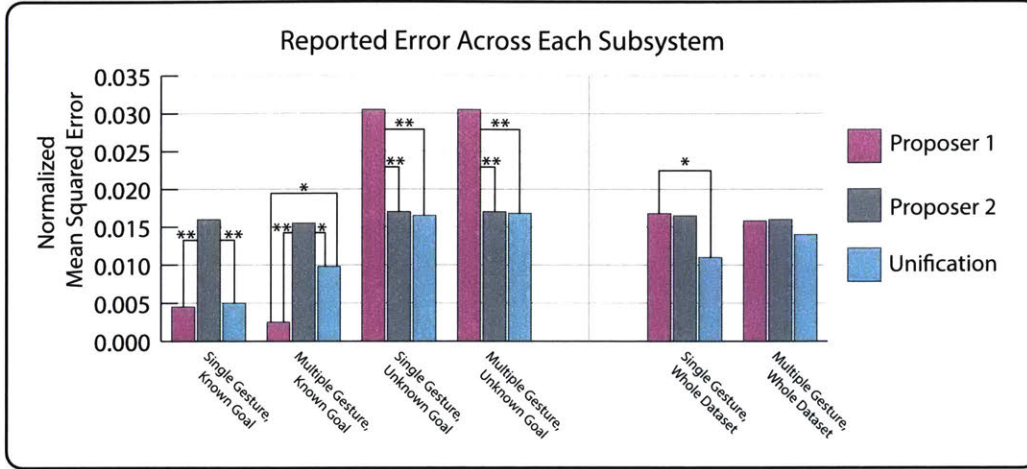


Figure 4-8: System performance against each test dataset. System-indexical: Proposer pump 1 returns predicted foregrounds  $F^P$  based on recognized objects. System-Saliency: Proposer pump 2 returns predicted foregrounds based primarily on the projection of deictic action on the scene itself,  $F^P$ . Competition: Resolving how to route each of the sub-pumps in the system to select the correct foreground. In this figure, \* refers to significance of  $p < 0.05$  while \*\* refers to significance of  $p < 0.01$ . Refer to Table 4.2 for further detail.

		Reported Significance		Average NMSE
		Pump 2	Unified	
SGKG				
SGKG	Pump 1	p<0.001	Insig.	0.004
	Pump 2		p<0.001	0.016
	Unified			0.005
		MGKG		
MGKG	Pump 1	p<0.001	p<0.03	0.003
	Pump 2		p<0.05	0.015
	Unified			0.010
		SGUG		
SGUG	Pump 1	p<0.001	p<0.001	0.030
	Pump 2		Insig.	0.017
	Unified			0.016
		MGUG		
MGUG	Pump 1	p<0.001	p<0.001	0.030
	Pump 2		Insig.	0.017
	Unified			0.017

Table 4.2: Reported significance and error of all subsystems with respect to the dataset. This table is separated into the dataset (by color) and the subsystem (Pump X + Unified). Significance values are reported using a Student's  $t$ -test, and the average normalized mean squared error is reported to three decimal places.



The previous section reports that indexical recognition systems are insufficient to resolve situated referential action alone and perhaps another system is needed to support such references. When the objects are known, indexical foreground prediction using high-level recognition of perceptual object boundaries performs better than saliency systems alone.

The findings show saliency-centered referencing onto foreground maps provides a better foreground estimation than indexical recognition systems where the trained object model is unknown to the robot. Figures 4-8 and 4.2 show that combining and building confidence values from the two proposal pump systems (Section 4.3.2) and resolving these through a novel deictic simulation system results in competitive foreground estimation across all situations.

The results from Figure 4-8 show that a competitive system (in blue) outperforms the the pump 2 proposals in all cases. The competitive system even performs better than both systems alone in combined datasets. The discrepancy between the known goal dataset and the indexical recognition system performance could be explained by the proposal pump 2 taking over when the recognition system should have overridden it.

## 4.6 Discussion

The findings suggest a model that leverages both saliency systems and recognition systems in competition to understand referencing behavior outperforms either system in isolation. This signal can be used for gating a learning mechanism or in other reflective contexts where the robot must indicate whether it recognizes the object at the point of reference. For instance, signals generated within the architecture during referencing behavior can be used both within the interaction and in crowdsourcing systems like that proposed in Chapter 2. One can envision a flexible robotic perception system that is not only able to follow gaze and deictic references (like those in previous joint attention systems) but can also use a new fixation point to understand if it should attempt to learn or reuse previously learned models. Lastly, this chapter contributes a model of self-as-simulator deictic action generation to further clarify the visual foreground that the robot extracts to enable visual learning from interaction in the future.

Using the estimated foregrounds, these foreground templates can be employed as exemplars for higher level learning systems to map internal deictic indexicals. These indexicals can be anchored into higher level reasoning systems through simple gaze fixation alone. The next chapter examines improvements on state-of-the-art, attention-centered neural networks that rely primarily on already-segmented (e.g., selected foreground) objects as input. This improvement is achieved by incorporating deictic action input to direct the visual learning system through a novel mechanism called a *deictic well*.

# References

- Admoni, H., Weng, T., Hayes, B., & Scassellati, B. (2016). Robot nonverbal behavior improves task performance in difficult collaborations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 51–58).
- Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *IEEE International Conference on Robot and Human Interactive Communication* (pp. 701–706).
- Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT Press.
- Blumberg, B. M. (1996). *Old tricks, new dogs: ethology and interactive creatures* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Brooks, A. G., & Breazeal, C. (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (pp. 297–304).
- Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1), 14–23.
- Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge University Press.
- Burke, R., Isla, D., Downie, M., Ivanov, Y., & Blumberg, B. (2001). *Creature smarts: The art and architecture of a virtual brain*.
- Butterworth, G., & Cochran, E. (1980). Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development*, 3(3), 253–272.
- Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *IEEE International Conference on Computer Vision and Pattern Recognition* (Vol. 1, pp. 886–893).

- DePalma, N., & Breazeal, C. (2015a). Object Discovery vs. Selection in Social Action: Benefits of a Competitive Attention System. In *8th International Workshop on Attention in Cognitive Systems (ISACS) at the International Conference on Intelligent Robots and Systems (IROS)*.
- DePalma, N., & Breazeal, C. (2015b). Sensorimotor Account of Attention Sharing in HRI: Survey and Metric. *2nd Annual Symposium on Artificial Intelligence and Human-Robot Interaction*.
- DePalma, N., & Breazeal, C. (2016b). Towards bootstrapping socially significant representations through robotic interaction alone: the joint guided search task. In *Fifth International Symposium on New Frontiers in Human-Robot Interaction at the Artificial Intelligence and Simulation of Behavior Convention (AISB)*.
- Doniec, M. W., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *6th IEEE-RAS International Conference on Humanoid Robots* (pp. 34–39).
- Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, 6(3), 205–254.
- Frintrop, S. (2006). *Vocus: A visual attention system for object detection and goal-directed search* (Vol. 3899). Springer.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580–587).
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.
- Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2), 135–169.
- Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence* (pp. 115–141). Springer.
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551.

- Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* (pp. 2204–2212).
- Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, 15(4), 211–229.
- Nagai, Y., & Rohlfsing, K. J. (2009). Computational analysis of motionese toward scaffolding robot action learning. *IEEE Transactions on Autonomous Mental Development*, 1(1), 44–54.
- Rolf, M., Hanheide, M., & Rohlfsing, K. J. (2009). Attention via synchrony: Making use of multimodal cues in social learning. *IEEE Transactions on Autonomous Mental Development*, 1(1), 55–67.
- Sauppé, A., & Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 342–349).
- Scassellati, B. (2001). *Foundations for a theory of mind for a humanoid robot* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Tomasello, M. (2000). *The cultural origins of human cognition*. Harvard University Press.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, 9(2), 125–147.
- Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press.
- Wolfe, J. M. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2), 202–238.
- Yu, C., & Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*.



## Chapter 5

# Reference following and object recognition with deictic wells

While the previous models of visual attention work by inhibiting entire regions of the scene and combining them piecewise into recognizable wholes, other models inspired by attention are proven efficient at recognition in sparse scenes as well. These artificial neural network models have the advantage of working on real world images (i.e., the NIST dataset) and have the capability to recognize objects against flat matte black backgrounds in no-clutter at high precision and accuracy comparable to deep convolutional networks (Mnih, Heess, & Graves, 2014; Ba, Mnih, & Kavukcuoglu, 2014). Additionally, these models are efficient, requiring orders of magnitude fewer instructions per second to emit classification labels from their architecture. While these models have yet to address critical issues in figure-ground selection, there is clear evidence that they have been successful at working with real world images. This chapter provides a single iterative step at moving these neurally inspired models toward a system that operates on real world images in a similar way to the SHARE system proposed in Chapter 4 and visualized in Figure 4-1. This iterative step takes social learning into account in a perceptual domain while integrating deictic action as input to be capitalized on in the neural architecture.

Humans have an innate and uncanny ability to efficiently direct their visual attention to obtain and extract relevant information in the scene while inhibiting other stimuli. Capitalizing on this and other social cognitive machinery, we can easily establish joint attention

with others as part of our communicatory competence. This work focuses on the use of nonverbal deictic gesture, such as pointing, gazing and intentional directive action, as a trigger to efficiently guide a social agent’s target search within a potentially rich visual scene. Deictic communication is defined to as an *action taken by an individual with the intention of directing another individual toward perceiving something not yet in the individual’s common ground*. Brooks (A. G. Brooks & Breazeal, 2006) and Clark (Clark, Schreuder, & Buttrick, 1983) share similar definitions along with the idea of Grice’s maxim of quantity (Grice, 1975). Less formally, they can be seen as *pointers* that must be resolved (A. G. Brooks & Breazeal, 2006; Admoni, Weng, Hayes, & Scassellati, 2016). In human-robot interaction (HRI), sharing compatible mental states between humans and robots at any one time is an ongoing challenge for researchers. Individuals (e.g., the robot or the human) must take synchronization actions by directing the perception system of one another towards some stimulus. The stimulus could either be novel or already known, but in the event that the stimulus is novel, the system must learn a new model to support later deictic referencing and priming. This work focuses on advancing state-of-the-art, deep neural attention networks that simulate saccade behavior in scenes toward learning and recognition. The results show the new model, rewarded by visual gesture following, simulates saccade behavior and maintains competitive recognition rates while improving the speed at which learning occurs. This chapter demonstrates improvements in learning speed over baseline RAM by trading off unrewarded explorations for rewarded deictic pointer following toward new stimuli.

To simulate autonomous saccade behavior, this work utilizes a rather recent development in deep neural networks that takes advantage of recognition by moving a small window intelligently across an image rather than using convolution across the entire image. This biomimetic approach (Bar-Cohen & Breazeal, 2003) is highly compatible with human-robot interaction that tends toward the design of *transparent* behavior that reveals the internal state of the robot through movements like saccades (Mutlu, Shiwa, Kanda, Ishiguro, & Hagita, 2009). Saccades are defined as the implicit behavior of an eye as it visually moves its focus of attention across an image or present scene. The rather recent insight in these models is to treat the extraction window as an *agent* that observes its state



and takes actions to move itself around the image more intelligently. This allows standard reinforcement learning algorithms to move the window efficiently to find and classify objects around its fixation point (or the moment of local saccade behavior). This work takes advantage of this innovation to bias the reward function in an attempt to direct the agent toward areas where the social partner is also attending. This reward innovation is called the *deictic well*. The deictic well utilizes observations of deictic gesture to resolve the pointer to its target. It also takes advantage of a reward function’s landscape to direct the learning agent toward more profitable states.

With the proposed approach, the behavior of the saccading agent is able to follow the deictic gesture provided by social partners while additionally learning about new visual stimuli faster than the baseline RAM model. Section 5.1 describes related work in utilizing small windows in deep recurrent models and also explores joint attention in human-robot interaction. Section 5.2 examines the architecture of the artificial neural network needed to produce this behavior while Section 5.3 describes the data collection and experimental setup. Encouraging results appear in Section 5.4 by comparing the proposed model against the baseline model examined in (Mnih et al., 2014). Finally, future work is discussed in Section 5.5.

### **5.0.1 Theory and hypothesis**

This work contributes a novel mechanism called a *deictic well* and makes two hypotheses:

- H1: The socially biased visual learning algorithm proposed in this work will improve classification accuracy.
- H2: The socially biased visual learning algorithm will speed up learning, guiding the agent’s observation window to the correct location.

The contribution extends a previous model of visual attention called *RAM* which is designed to reduce the number of observations within an image to classify and recognize the object in the scene. It has comparable accuracy to competitive convolutional networks (LeCun et al., 1989). The outcome of these hypotheses could have significant impact on

situated perception learning agents that must adapt to perceive new objects in new environments.

## 5.1 Related Work

### 5.1.1 Deep recurrent models of attention

Deep recurrent models of attention are powerful tools that mimic biological visual attention to perform visual search (Wolfe, 1994) tasks within images. Computer scientists use them to:

- Perform efficient classification and object recognition tasks on large images (Mnih et al., 2014),
- Do multi-object detection tasks (Ba et al., 2014),
- Draw learned primal sketches (Gregor, Danihelka, Graves, Rezende, & Wierstra, 2015),
- Learn to perform addition after finding digits in an image (Ba et al., 2014),
- And even generate image captions (Xu et al., 2015).

The core innovation in attention-based systems is to direct the resources of the algorithm toward a specific region to identify objects at that position, as opposed to trying to find a single object in the scene (object detection). In effect and at its most abstract, it is a problem of computing which object,  $O$ , is at location  $l$  using  $p(O|l)$  instead of finding the location of the object,  $l_o$ , using  $p(l_o|O)$ . Recurrent neural networks offer a powerful way to direct the resources of the network toward taking the most informative observations to classify an object under a given location. This work attempts to augment previous models to accept deictic action observations.

### 5.1.2 Joint attention in agent learning

Many evolutionary psychologists, social roboticists and synthetic epigenetic scientists have focused on the emergence of one early communicatory competence mechanism called *joint attention* (Tomasello, 1995). It is considered the earliest form of communication to emerge during development and seems to be present only in humans and the great apes (Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998). Computer scientists have proposed computational approaches to modeling the bio-inspired joint attention behavior. Triesch et. al. (Triesch, Teuscher, Deák, & Carlson, 2006) developed a computational model of gaze following behavior in infants. This involved reinforcement learning in a small discrete world to learn to follow gaze to objects. Schmidhuber and Huber (Schmidhuber & Huber, 1991) studied artificial fovea trajectory for moving target detection and Ognibene et. al. (Ognibene & Baldassare, 2015) suggested a fovea based recognition model using reinforcement learning. While these are impressive systems, the SHARE system seeks to make progress toward learning a perceptual label from social signals, in essence working toward taking advantage of synchronicity of different stimuli within the interaction (Rolf, Hanheide, & Rohlfing, 2009).

This chapter introduces a multi-agent learning algorithm that is integrated with a robotic perception system to jointly learn percept names and follow deictic action toward new targets. Deictic observations are generated synthetically from the data collected in Chapter 4 that takes advantage of observations made from a human-robot dyad. These actions are used as input along with a novel database for the learning algorithm to gain information about targets in the scene. Future work on this project toward integrating the perception system back into an embodied and interactive social learning system is discussed in Section 5.5.

## 5.2 Model design

The problem this chapter seeks to address is introducing reference following (whether from gaze or gesture) to state-of-the-art neural network models. The model presented is an extension of the work of Mnih in which he used recurrent neural networks (RNNs) (Mnih et

al., 2014) to control a simple observation window. At present, recurrent attention models are agents interacting strictly with the environment and not with social others. The model itself is called a Recurrent Attention Model or *RAM*. This simple window simulates foveal saccades around an image seeking to identify objects. While these deep learning architectures have been shown to be very successful in various visual perception tasks, they are not augmented to be socially biased. For a synthetic attention system to become a synthetic *joint* attention system, it must make use of the deictic signal proposed earlier in this chapter. Yu and Smith (Yu & Smith, 2016) have found evidence that attention is directed by gaze just as much as it is by gestural feedback. One theoretical unifying representation for these types of input is the deictic code (Ballard, Hayhoe, Pook, & Rao, 1997; A. G. Brooks & Breazeal, 2006) or internally and externally represented pointers that are maintained and followed by the cognitive processes. The dRAM stands for *deictically stimulated Recurrent Attention Model*, extending Mnih’s model to support learning to follow referential gesture in a perceptual domain. Section 5.2.1 discusses the previous visual attention model and Section 5.2.2 discusses the modification to the model that leads to deictic following.

### 5.2.1 The baseline RAM model

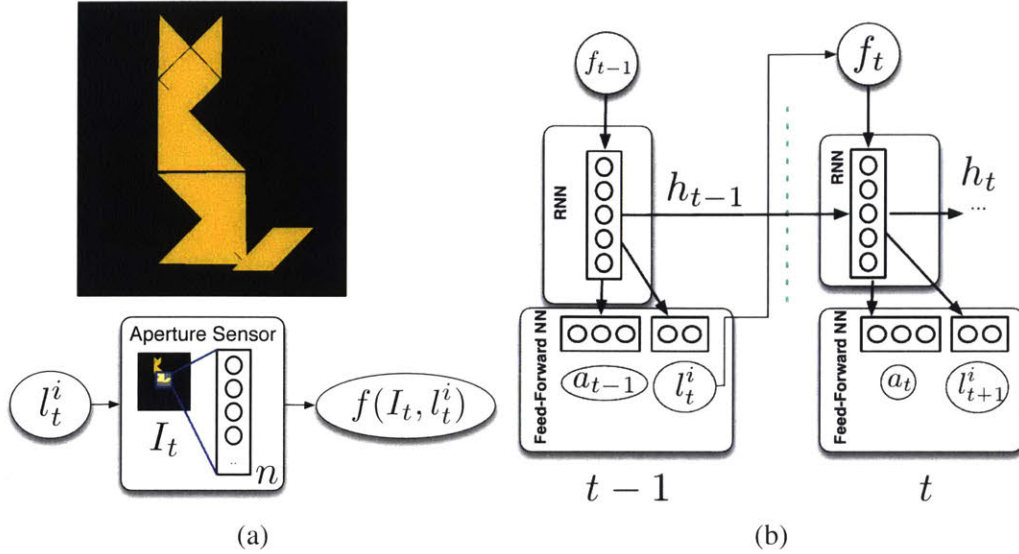


Figure 5-1: Summary of the RAM model. (a) An example taken from the crowdsourced dataset that resembles and is labeled “cat” (top). The basic extraction network uses the previously computed location position,  $l_t$ , and the image data at time,  $I_t$  (bottom). (b) The observed state,  $f_t$ , is used to train the hidden recurrent network which in turn predicts a reward, then chooses the next best action to take and the next location for moving the artificial fovea.

The recurrent attention model (RAM) uses a sensor (see Figure 5-1a) that extracts a small image patch at location  $l_t$  on an image  $I_t$  at time  $t$  which is additionally augmented with the location,  $f(I_t, l_t) = [I_t l_t]$ , a state vector. Foveal location as marked using the  $i$  superscript,  $l_t^i$ . The feature vector is used as an input to a recurrent neural network,  $h_t$ , which is used to predict the reward that it would receive after  $t$  action steps. The action space is an augmented space where it includes both actions to change the location  $l_{t+1}^i$  as well as an emission labeling action,  $a_t$  which simply emits an ordinal or a *no label* action.

#### 5.2.1.1 Reinforcement learning

Figure 5-1 visualizes just the reinforcement learning agent. In this figure, the observation itself is augmented with the previous state to make up  $h_t$ . In standard reinforcement learning texts, basic policies are defined as actions taken at specific states,  $\pi(s) \rightarrow a$ . In this case, the internal state is used to emit the classification action, predict the reward and the

next location,  $\pi(h_t) \rightarrow [a_{t+1}, l_{t+1}^i, r_{t+1}^p]$ . A single layer is trained to map the outputs from the internal state,  $h_t$  to the desired outputs during learning.

The RAM model emits classification actions for each time step but in the case of this model, only the time step at time  $T$  is considered. This means every test image will be tested with  $T$  observations and will output  $a_T$ , resulting in a single classification action for an image. The learning process is described further in Section 5.2.3.

### 5.2.1.2 Reward signal

The architecture leverages a reward signal during training when it has chosen the correct action after  $T$  steps. The reward function used in RAM depends on the sequence of actions taken on the image itself.  $R_T = \sum_{t=0}^T r_t$ . Since the agent is trained as a reinforcement learning agent to follow the gradient and maximize its reward, it is sensitive to changes in the reward function gradient itself. The dRAM model presented next takes advantage of this mechanism to teach the system to reward gesture following behavior of the agent when it is observed toward both learning and classification.

## 5.2.2 The extraction window and deictic observation

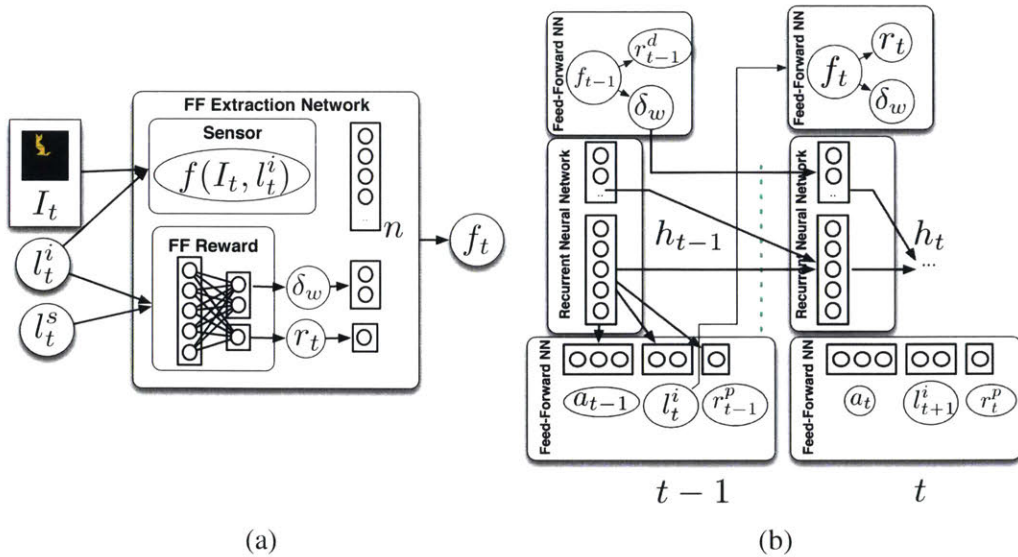


Figure 5-2: The proposed dRAM model. (a) Basic extraction that additionally computes the gradient  $\delta_w$  and a reward  $r_t$  toward an observed, social deictic pointer,  $l_t^s$ , at time  $t$ . (b) The modified architecture that takes the gradient into account in its hidden state  $h_t$ .

The deictic RAM (dRAM) model incorporates deictic indexing into the learning process (see Figure 5-2). This work introduces the concept of the *deictic well*. In previous models, state is estimated from observations made from the extraction network, referred to as a *glimpse*. A special treatment is made to learn an action policy based entirely on rewards observed when the recurrent network takes the correct classification action in the right location. The insight of the *deictic well* is to take advantage of a social deictic action to induce a gradient that leads the learning agent toward discovering the maximal reward at the correct location. Consequently, the action policy should learn to utilize the deictic gradient features toward the referenced point to follow the pointing gesture toward the associated classification action. This section will cover the needed updates to the RAM model to support deictic learning through deictic referencing.

The dRAM model capitalizes on a basic extraction window to pull features out of the image for recognition (see Figure 5-2). The glimpse observation network is modified to take as input an image,  $I_t$ , and a previous location  $l_t^i$  at a time  $t$  from its previous action to output a feature vector,  $f(I_t, l_t^i)$ . The extension presented in this chapter focuses on grading the reward function using the current location and the observation of a social deictic action's intended location  $l_t^s$  as parameters. Another way to think about  $l_t^s$  is to imagine that this location is the center of the deictic well's reward. The contribution to the reward function of the deictic well is defined to be the normalized  $\mathcal{L}_1$ -norm,  $r_t^d = \frac{1}{c} \sum_m |l_t^i - l_t^s|$  (see Figure 5-3 for a profile) where  $c = w * h$ , is the product of the width and height of the image. The resulting reward function used to train the network is then defined as:

$$R_T = \sum_{i=0}^T r_t^p + r_t^d \quad (5.1)$$

$$R_T = \sum_{i=0}^T \left( r_t^p + \frac{1}{c} \sum_m |l_t^i - l_t^s| \right) \quad (5.2)$$

Additionally, the gradient features of the deictic well,  $\delta_w = |l_t^i - l_t^s| \in \mathcal{R}^2$  are provided as features (see Figure 5-2),  $f_t = [f(I_t, l_t^i), \delta_w, r_t]$ , to the state,  $h_t$ . The resulting feature set,  $f_t$  is composed of the 131 units in which 128 units are the current output of the window's



observation 2 and units represent the deictic well's gradient and magnitude. The final unit is the reward prediction computed from the difference between the current locus of attention and the other agent's directed action target which is the intended change of the internal extraction networks' location.

### 5.2.3 Training the network

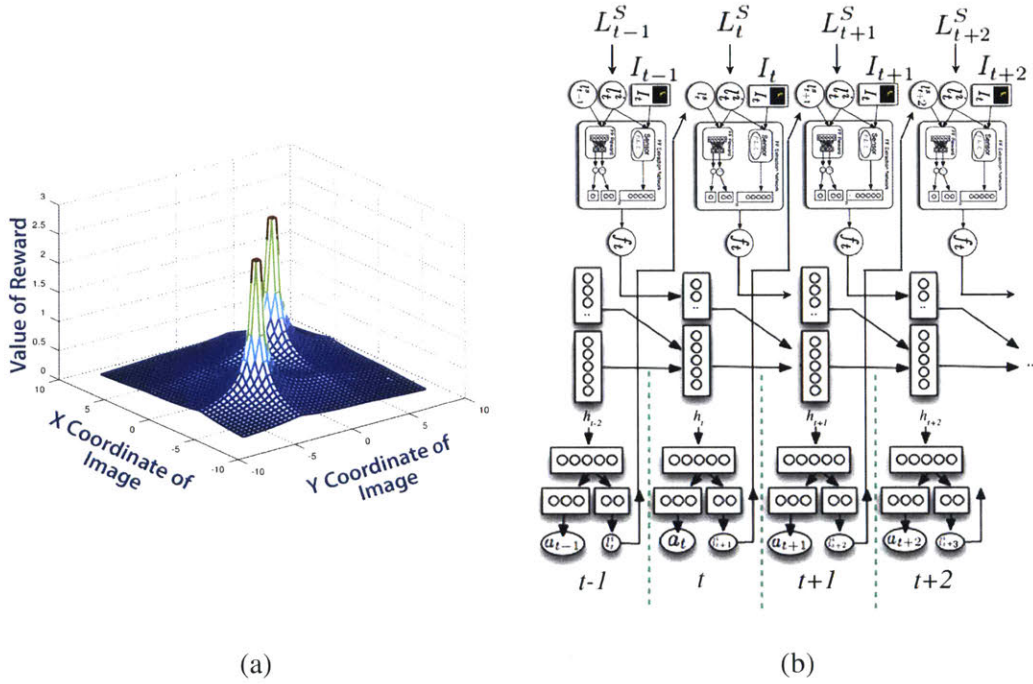


Figure 5-3: Full dRAM architecture with deictic sampling. (a) Profile of a deictic well (the component that is summed with the previous reward function), (b) the full artificial neural architecture.  $l_t^t$  is sampled uniformly from a set of pointing gestures collected from human participants interacting with the robot ( $L_t^S$ ).

A time-indexed recurrent network with 256 hidden units is trained separately using  $f_t$  as input to the hidden units state at time  $h_{t+1}$  (see Figure 5-3). Learning in neural networks typically uses back-propagation to perform weight updates between layers of artificial neurons. In the case of recurrent neural networks (RNNs) which are connected to themselves for sequence learning, the loop is unrolled for  $\rho$  steps and treated as a standard multi-layer perceptron. This process is called back-propagation through time or *BPTT* (Werbos, 1990). These experiment sets unroll the loop to unroll over 7 time steps ( $\rho = 7$ ). Back-propagation

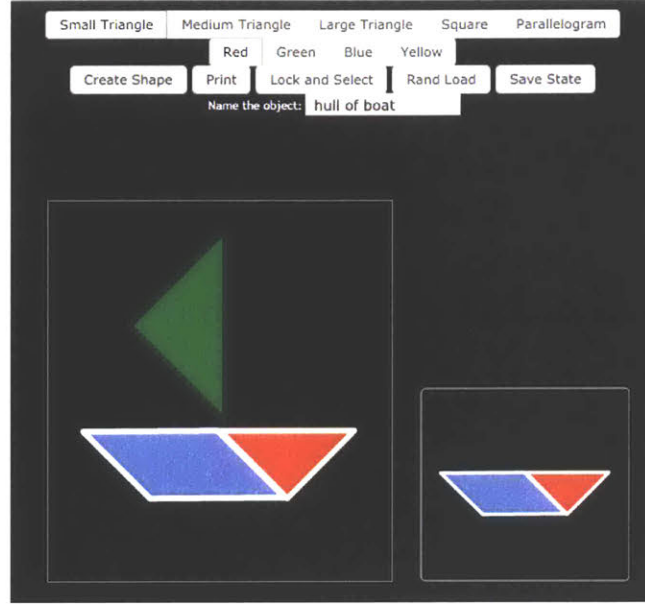


is a process in which the inputs are propagated through the network to the end of the network where they are checked against an output. Error is computed (in this experiment's case, the dRAM network uses negative log likelihood criterion ( $loss(a_t, x) = -a_t[x]$  where  $x$  is the correct classification action and is constant for a particular image). Standard BPTT can be very problematic for networks that must make parameter updates for actions taken each time step of the loop. Following closely the work of (Mnih et al., 2014), the REINFORCE algorithm is used to estimate the gradient,  $J$ , at each loop step of the RNN. In this case, a high-dimensional estimate of the gradient is required for all parameters of the network:  $\theta = [\theta_e, \theta_r, \theta_a]$  representing the parameters of the extraction network,  $\theta_e$ , the parameters of recurrent network,  $\theta_r$ , and the parameters of the action network,  $\theta_a$ . REINFORCE (R. J. Williams, 1992) notes that to train the network as a reinforcement learning agent, it must maximize the reward across the  $T$  action steps prior to the classification action:  $J(\theta) = E_{h_{1:N}}[R]$  where  $h_{1:N}$  is the sequence of  $T$  states observed during trained policy at the present epoch. Mnih (Mnih et al., 2014) shows a sample approximation of the gradient needed to supply to the BPTT to update the weights in the following:

$$\nabla_{\theta} J = \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^T \nabla_{\theta} \log \pi([l_t^i, a_t | h_{1:t}; \theta) R^i \quad (5.3)$$

The expectation of the gradient is computed for the current agent,  $\pi$ , for  $i = 1 \dots M$  episodes (within the epoch of training) and  $[l_t^i, a_t]$  are the interaction sequences. REINFORCE adjusts the parameters  $\theta$  so the log-probability of the chosen actions that led to a higher cumulative reward is increased. Overall, this procedure maximizes the cumulative reward of the state-action trajectory and the reward provided by taking the correct classification action. If the reader is interested further, they may refer to (R. J. Williams, 1992) for a detailed explanation. The algorithm is used with few changes to train the network and no changes are required of the training method itself. The process of walking through state-action saccades makes recurrent attention networks very powerful and efficient object detectors while preserving pro-social interaction behavior.

Finally, BPTT is used to propagate the reward gradient across the recurrent network



(a)



(b)

Figure 5-4: Data Collection. (a) The crowdsourced data collection system. On-line participants can construct tangrams using a mouse and keyboard (left). Highlighted sections (right) are used to select the foreground component and a label can be sourced on-line (e.g., “hull of boat” in this example). This allows us to collect subfigure labels for many different tangram figures as well as full figures. (b) A few sample tangrams collected from the dataset online.

and update the weights of the embedded RNN. The main inputs to the system are the label, the image and the observation of a deictic action that is meant to direct the attention of the perception system toward a region of the image that may be more profitable. Reward gleaned from both the classification action and the deictic well is summed together during back-propagation training of the neural network where  $r_t^d \ll r_t^p$ .

### 5.3 Data collection and experiment

To test the hypothesis (H1) that accuracy improves with social referencing and this model can learn to follow gesture toward recognition (H2), this model was tested with a dataset collected from participants on-line comprised of objects, figures and labels collected from NIMBUS. The interface designed for NIMBUS was built to crowdsource tangram figures<sup>1</sup> (see Figure 5-4a). The interface itself has the ability to construct tangrams, select subparts and pixel maps while providing labels for the selected regions. Next, the dataset was consolidated and full figure images were rendered (labels applied to the entire tangram figure). Then the figures not properly constructed (tangram pieces strewn about) were cleaned from the dataset. A sample from the tangram dataset can be viewed in full-color rendered form in Figure 5-4 and also processed in grayscale for use in the learning algorithm in Figure 5-6a. Figure 5-7a shows H1 tested by measuring classification accuracy and H2 tested by analyzing the trajectory after just 100 time steps on the image.

To generate the deictic actions  $l_t^S$  from the dataset, tangram figures were separated into subfigures and re-rendered for each subfigure, along with the associated labels. Next the set of deictic gestures were aligned,  $L^S$  (i.e., social referencing locations) to the images themselves. While many types of referential gestures were observed in Chapter 3, the data used in this experiment and which make up  $L^S$  only analyzes single action, simple deictic-conic representations for this experiment.

The dataset included a total of 43 tangram scenes broken down into 125 subfigures to make a total of 168 possible reference labels to figures with an average 7.29 actions per figure ( $\bar{l}_t^S = 7.29, \sigma = 1.3$ ). This set of actions made up the discrete distribution of deictic actions used for sampling from each scene during learning.

For the experiment, the set of actions for each figure were sampled uniformly:

$$L_t^S \sim [l_t^s]$$

The architecture used the actions as input to the system for training (Figure 5-3b). Six experimental groups were defined, dRAM<sub>1-6</sub>, where the subscript refers to probability

---

<sup>1</sup>Available at: <http://nimbus.media.mit.edu>

that the network would successfully observe a referencing action (e.g., dRAM<sub>1</sub> that has a probability of 0.1 of observing a deictic action from the dataset, dRAM<sub>2</sub> at a probability 0.2, etc.). Finally, the baseline case (RAM) was defined to be the model used by (Mnih et al., 2014) that did not take advantage of deictic action to direct its learning.

## 5.4 Results and discussion

The following metrics were used to understand the effect of deictic action on perceptual learning: elapsed time, classification and glimpse location. The time elapsed at which learning occurred in the baseline RAM model was measured and then compared against learning at various sampling probabilities.

### 5.4.1 Elapsed time in epochs

To understand the utility of current forms of RAMs against this model, the average duration of learning was measured with the tangram dataset. The experimental model was tested on both a quad core Intel i7 processor which observed each epoch to take approximately 1.5 seconds ( $\bar{x} = 1.44s, \sigma = 0.28s, n = 344$ ) and on a 16 core Intel Xeon, observed to require 0.1 seconds per epoch ( $\bar{x} = 0.14s, \sigma = 0.02s, n = 205$ ). It is clear that over the course of exposing the learning algorithm to a particular stimuli (the image) within an interaction scenario, timing was critical. The fewer epochs it took the algorithm to converge to a particular approximate solution, the better the algorithm could be within the same period of time. The main objective was to preserve recall while drawing the learning agent to the correct stimulus and to do this as fast as possible, thus making duration of an epoch incredibly important to the interaction system. A fixation time of 20 seconds (200 epochs) in an interaction could be considered a breakdown for the user's experience.

### 5.4.2 Deictics speed up convergence

Deictic action was observed for each scene in the collected dataset and learning was simulated using deictic reference and images as input to the algorithm. For each tangram, the

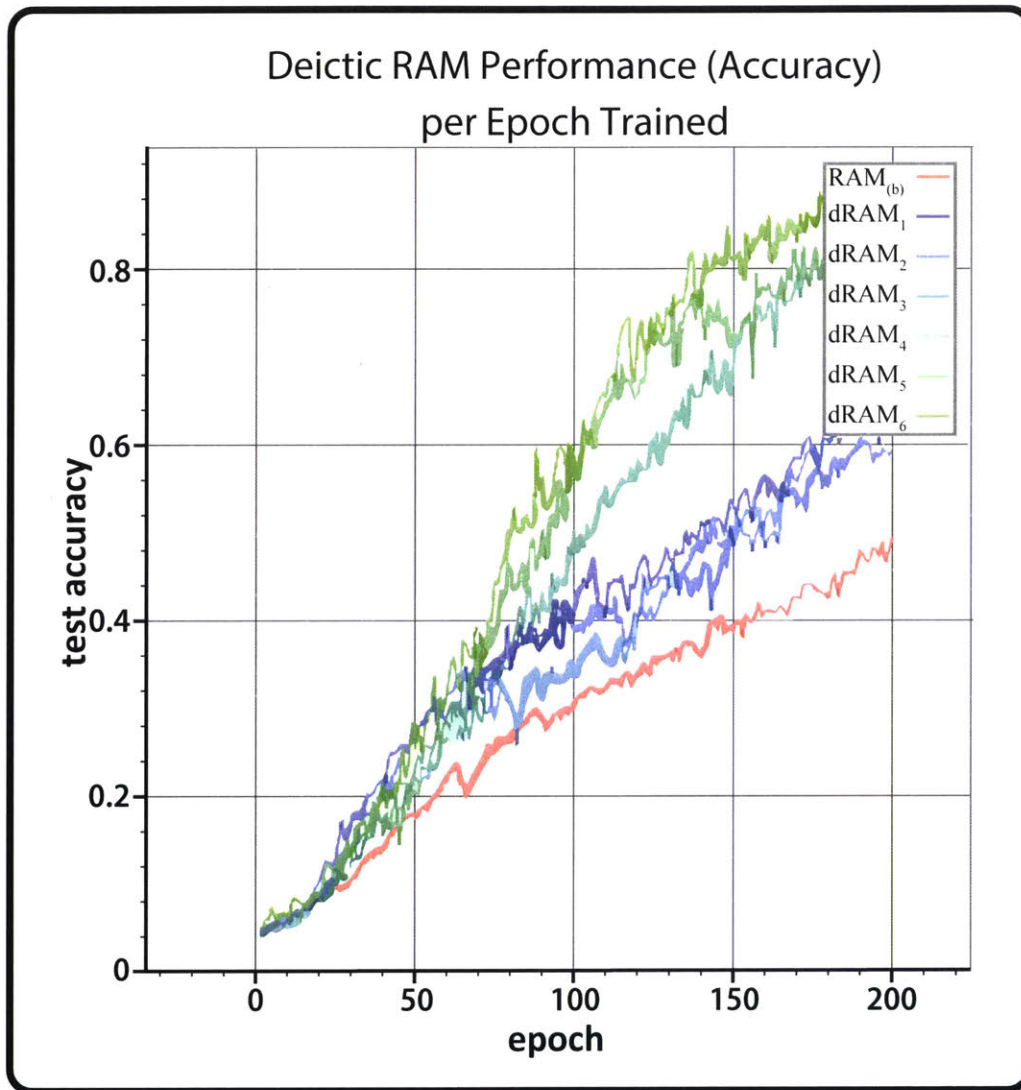


Figure 5-5: Test accuracy progress over each epoch plotted with 7 groups. The baseline RAM model converges at a slower rate than the extended model, dRAM. Each dRAM subscript represents the probability of observing a deictic action from the dataset. It is clear that more social interaction has a positive effect on this model.

deictic actions were randomly sampled with an associated given probability (i.e.,  $p = 0.1$  for group dRAM<sub>1</sub>,  $p = 0.2$  for group dRAM<sub>2</sub>, etc.). For each group, the model simulated learning across 200 epochs for 10 runs. Figure 5-5 describes the average accuracy for each epoch as the network modifies its weights to learn to map a specific label to the stimulus. It is clear from Figure 5-5 that after initial burn-in, each incremental increase in the probability of observing deictic action improves the speed at which the learning algorithm converges.

This model extends work by (Mnih et al., 2014) but with a focus on its applicability to learning interactively with a participant. Findings show the faster an agent can learn to reliably map a sign to an image at a spatial position, the more actions may become available to the agent, based on a short interaction with the environment and a social partner. Learning interactively will require significantly more improvements in speed prior to its applicability in interactive situated robotics. Next, this chapter will analyze the model's ability to follow deictic action.

### 5.4.3 Learned policy performance

The process of the agent saccading on the image is simulated for the classification task to better understand how this model of socially directed selective attention performs on a given image. Figures 5-6 and 5-7 show the performance of the baseline RAM model against the dRAM model, respectively. The figures show 10 randomly chosen objects (the x-axis) from the dataset with the position of the fixation point (the box) after 100 observations (randomly chosen). The RAM model shows the results of the agent's fixation location from each of the 10 trained models described above (y-axis). The dRAM model shows the results from the final learned model. It is notable that after just a few (100) observations, the dRAM model is primarily on task, making observations on the stimulus itself. Mnih et. al. point out that by having these deep recurrent models, the recurrence itself can be thought of as an agent making partially observable and highly stochastic observations about the world. Since the agent itself can be directed toward other stimuli, it can be *distracted* from the current task to move into more profitable spaces to classify, thus making this a

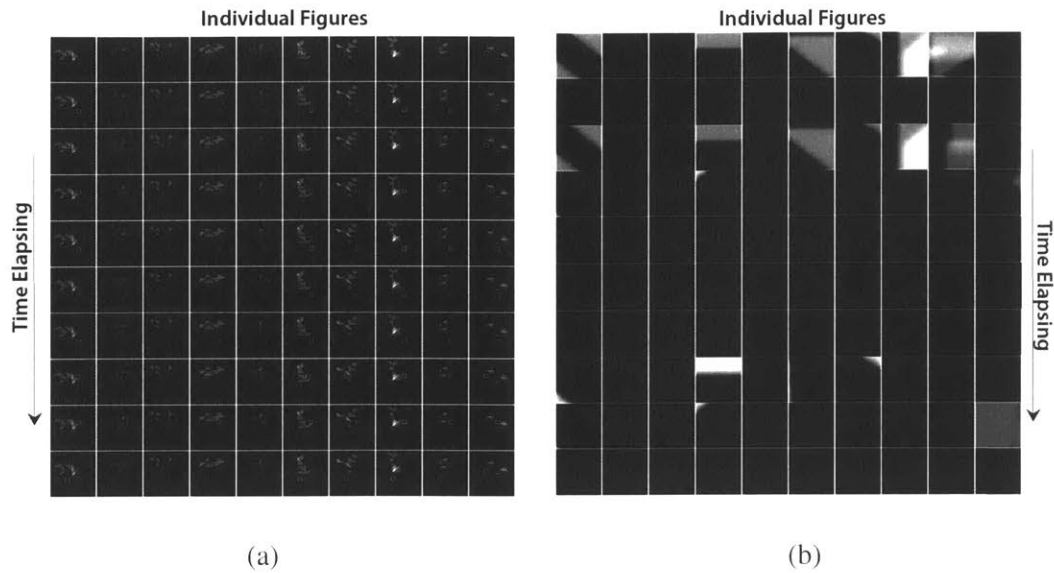


Figure 5-6: Performance of the RAM Model after 100 observation steps. Vertical axis is the model's performance from the 10 test simulations, each row being a particular simulation. (a) The given scene with the fixation point overlaid on top, (b) The extracted patch used by the algorithm. Discussion of these results can be found in Section 5.4.3.

pro-social classification model as well.

## 5.5 Summary

While gesture recognition can be seen as a classification problem, gesture understanding involves resolving the sign or reference to an underlying concept. This process is called anaphoric or exophoric reference resolution in the computational literature. This insight is primarily inspired by human-human social learning in which people exchange gestures and follow gaze as well as pointing gestures to target areas in the scene. In this case, the gesture (sign) is mapped to a visual stimulus (tangram figure) and an associated label is used to additionally map words (lexical signs) to the same stimulus. This particular system falls into a category of artifact that is attempting to map signals to image data, in effect learning to follow deictic action to its perceptual center.

The goal is to extend this work in many ways: one is to test this system with real human participants to perform interactive percept learning on the robot. This effort will involve smoothing saccade behavior for motor control and incorporating aspects of memory into an

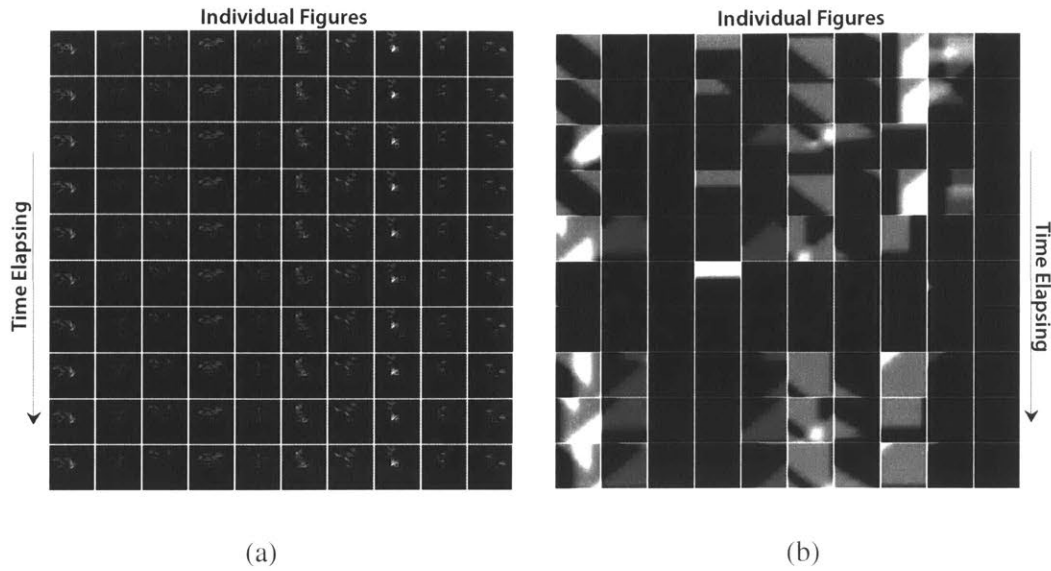


Figure 5-7: Performance of the dRAM Model after 100 observation steps. Vertical axis is the 10 learned models,  $dRAM_{1-10}$ . (a) The given scene with the fixation point overlaid on top. (b) The extracted patch used by the algorithm. Discussion of these results can be found in Section 5.4.3. Note that the observed patch is overwhelmingly sampling from the figure itself after 100 observations in the dRAM model vs. the baseline RAM model (Figure 5-6).

interactive system that must perform label and reward attribution. Additionally, this system is intended to be integrated with an early model of foreground reasoning that will allow this system to choose the underlying stimuli for use in learning to a mapping. This model is intended to be extended to map robotic behavior to the simulated saccade behavior. It will also generate deictic action (pointing, sweeping and palm gestures) that allows the robot to point and gesture in many of the ways that people do. Additionally, future work should address extending some of the previous work from Chapter 4 that predicts the foreground of the human participant and takes actions toward synchronizing it with the robot's goal.



# References

- Admoni, H., Weng, T., Hayes, B., & Scassellati, B. (2016). Robot nonverbal behavior improves task performance in difficult collaborations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 51–58).
- Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.
- Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(04), 723–742.
- Bar-Cohen, Y., & Breazeal, C. (2003). Biologically inspired intelligent robots. In *Smart Structures and Materials* (pp. 14–20).
- Brooks, A. G., & Breazeal, C. (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (pp. 297–304).
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.
- Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22(2), 245–258.
- Gregor, K., Danihelka, I., Graves, A., Rezende, D., & Wierstra, D. (2015). DRAW: A Recurrent Neural Network For Image Generation. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)* (pp. 1462–1471).
- Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics, Vol. 3, Speech Acts*. ed. by Peter Cole and Jerry L. Morgan, 41–58.

- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4), 541–551.
- Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* (pp. 2204–2212).
- Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (pp. 61–68).
- Ognibene, D., & Baldassare, G. (2015). Ecological active vision: Four bioinspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Transactions on Autonomous Mental Development*, 7(1), 3–25.
- Rolf, M., Hanheide, M., & Rohlfing, K. J. (2009). Attention via synchrony: Making use of multimodal cues in social learning. *IEEE Transactions on Autonomous Mental Development*, 1(1), 55–67.
- Schmidhuber, J., & Huber, R. (1991). Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 2(01n02), 125–134.
- Tomasello, M. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, 103–130.
- Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, 9(2), 125–147.
- Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10), 1550–1560.
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4), 229–256.
- Wolfe, J. M. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2), 202–238.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., ... Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv*

*preprint arXiv:1502.03044*, 2(3), 5.

Yu, C., & Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*.



## **Chapter 6**

# **Sustainability of longitudinal human-robot joint attention in dynamic environments: adaptation and imprinting**

Joint attention is a critical problem that shows up in many applications of human-robot interaction (HRI). The phenomenon underpins developmental and social learning (Nagai, Hosoda, Morita, & Asada, 2003; Doniec, Sun, & Scassellati, 2006), collaboration (Yu, Scheutz, & Schermerhorn, 2010) and, more importantly, makes up a core cognitive mechanism underpinning dialogue and much of social cognition in human peer groups (Kaplan & Hafner, 2006; Seemann, 2011; Tomasello, 1995, 2000; Carpenter, Nagell, Tomasello, Butterworth, & Moore, 1998). Joint attention may also play a critical factor in the new and burgeoning field of socially assistive robots designed to change behavior in response to stimuli (Matarić & Scassellati, 2016). While researchers still need to deeply model and understand the phenomenon of attention in order to build synthetic systems (see (Tsotsos, 2011; Bridewell & Bello, 2015) for a sample of the types of progress being made), they have seen success with attention's behavioral outcomes on social agents. Researchers and engineers can also mimic simple behavioral effects through behavior-based control and

some forms of agent learning. Unfortunately, many of these algorithms focus primarily on static environments and researchers often encounter trouble generalizing to dynamic environments. As an attempt to move forward, this study focuses on an investigation of the engagement and behavioral effects of longitudinal joint attention from human-robot interaction in dynamic scenes.

Human-robot interaction systems focus primarily on the outward behavior of the robotic architecture and its effects on a social peer. Due to this focus on behavior, joint attention is now understood as the behavioral manifestation of an underlying cognitively inspired robotic subsystem responsible for the mimicry of a subset of human behaviors that lead to joint attention in humans. This definition leads one to focus primarily on the behavior of referential gesture and gaze following through multiple sensory pathways (Yu & Smith, 2016). The phenomenon of human-robot joint attention is one of the major reproducible experiments that occurs in the study of HRI. As a result, there are a number of successfully reproduced experiments in bidirectional deictic gesturing toward gaze following (Sauppé & Mutlu, 2014). Despite this progress, building large cognitive architectures around the emergence of joint attention has not received the level of attention it deserves (see (Kaplan & Hafner, 2006)). The focus of this contribution is not on autonomous architecture but on the interaction and behavioral effects of longitudinal joint attention on human participants, achieved by biasing core aspects of the architecture (described in more detail in Section 6.2). This research emphasizes the following assertion: a cognitive system that is built around dynamic, exchanged social attention should stand the test of time and remain coupled throughout an interaction. The objective is to understand more deeply the types of interaction effects that can be expected.

### **6.0.1 Theory and hypothesis**

In human-robot interaction, robots are designed to mimic human-like behavior in order to understand its impact on interaction and behavior. Some examples of these effects are to ease the human partner's cognitive load (Jung et al., 2013) and improve the legibility of robot motions (Dragan, Lee, & Srinivasa, 2013). Roboticists frequently make assumptions

regarding interactive artificial agents that must include specific observations within their environment that involve an interaction partner. These hypotheses frequently don't weigh the impact of ignoring deictic gesturing on the sustainability of the joint coupling. The hypotheses here are based on past research in joint attention in which *motionese* is shown to play a major role in how caregivers draw attention to certain objects in the environment by incorporating dynamics into attention models. Inspiration is also derived from single agent-environment attention models that combine saliency and visual search within a visually guided situated robotic architecture. Finally, observations are leveraged from psychology literature to generate hypotheses around the effects of selective attention and attention's sister problem, memory, in how participants observe and remember events.

Dynamic environments provide many potential motion moments and distractors that an agent may attend to as well as a new dimension for the problem of joint attention: time. Time shows up in dynamic environments where objects and subjects are moving within the scene and also in longitudinal settings where a human peer has extended exposure to the experiment. In the following case, the experiment lasts for approximately 40 minutes. The following study is designed around the following hypotheses:

- H1: Human participants follow a robot's gesture toward shared dynamic scenes.
  - This hypothesis follows from two questions frequently on the minds of human-robot interaction scientists. Despite robots being machines, should human partners treat them as intentional agents and follow *their* directional social cues? Does previous research in human-robot interaction in static scenes also follow into dynamic scenes?
- H2: Human participants will follow socially-directed action (deictic gesture) over other best practice competing models of attention like saliency maps.
  - An agent that only interacts with the environment primarily leverages saliency maps to direct its attention.
- H3: A human participant will correct a robot that is distracted by other stimuli and bring its gaze back to the shared scene.

- H4: A robot’s deictic gesture can direct a human participant’s gaze toward low salience events in a dynamic scene and the directed event can also imprint itself in human memory.
- Following H1 & H2, if the robot successfully directs the human participant to a specific location, does the human participant remember a lower salience event they are directed toward?
- H5: Joint attention with a socially contingent robot is sustained throughout the entire interaction.
- Focusing and holding attention between two agents can be a challenge. The hypothesis in this case is throughout the entire interaction, participants will stay engaged and jointly couple their gestures with the robot’s gaze and gesture if the robot is perceived as socially contingent.

These hypotheses represent a subset of questions around joint attention in dynamic scenes where gaze and gesture are tracked in a tight loop with the robot.

## 6.1 Related work

Scasselatti (Scasselatti, 2001) presents some of the first work in gaze following, an early predecessor to joint attention. Some of these systems moved from behavior-based control to learning to follow gaze. Nagai (Nagai et al., 2003) and Triesch (Triesch, Teuscher, Deák, & Carlson, 2006) built systems to learn to follow gaze using neural networks and reinforcement learning, respectively. Following this seminal work, the bidirectionality of gesturing back to a social partner became the emphasis of subsequent systems. Doniec (Doniec et al., 2006) presents work on using active learning to map referential gesture or gaze to objects in the world. While this work is critical to moving forward with joint attention, the research documented here moved from learning to follow gaze in static environments back to behavior control in dynamic scenes where objects move fast and can become occluded



quickly. The system presented next uses the idea of a joint saliency map built with a social partner through the exchange of gesture.

### **6.1.1 Deictics in robotics**

Referential gesture that is dependent on context is sometimes referred to as deictic. Referential gesture and deictic use in human-robot interaction has been studied in various capacities. Most recently, Admoni and Scasselatti show that the use of deictic reference improves performance in difficult collaborations (Admoni, Weng, Hayes, & Scasselatti, 2016). Brooks and Breazeal (A. G. Brooks & Breazeal, 2006) present a model of multi-modal deictic use in communication that leverages grammar models to generate deictic gesture. Sauppé and Mutlu (Sauppé & Mutlu, 2014) present work on specifying a number of categories of deictic use in reference which include pointing, presenting, touching, exhibiting, grouping and sweeping. Thomaz et. al. (Thomaz, Berlin, & Breazeal, 2005) additionally demonstrate a model of shared attention for a social referencing task where affect is communicated by information seeking actions toward caregivers about objects in the shared frame of attention. Berlin, et. al. (Berlin, Breazeal, & Chao, 2008) show other social cues associated with proxemics may contribute to intentionally passive deictic use in a social learning context. Other effects such as synchrony (Rolf, Hanheide, & Rohlfling, 2009) and motionese (Nagai & Rohlfling, 2009) may also contribute to the intentional capitalization of innate biases of attention if they are jointly observed between the agents. This work extends previous findings into dynamic scenes and narrative shared between naive human participants and a robot.

### **6.1.2 Visual attention in robotics and vision**

Researchers mainly use saliency models to address visual attention in robotics. VOCUS (Frintrop, 2006) was a major attempt to unify a number of attentional phenomena including goal-directed visual search, learning about objects via saliency maps and a curiosity system driven to find new, novel objects in its world. VOCUS did not include a model of human partners and did not address joint saliency in a similar fashion to past work but it was a

landmark system for robot learning through attention alone. Meanwhile, vision researchers have also explored visual attention for everyday images. Itti, Koch, and Niebur's Neuro-Vision Toolkit (NVT) (Itti, Koch, & Niebur, 1998) led to many more models of saliency. NVT is still state-of-the-art in terms of generating saliency maps from images that represent accurate estimation of the likelihood that subjects will fixate at specific points of an arbitrary image (Hou, Harel, & Koch, 2012; Harel, Koch, & Perona, 2006). Finally, Tsotsos' work (Tsotsos, 2011) presents a model of visual search inspired by Gelade and Triesman's leakage model (A. M. Treisman & Gelade, 1980). However, other researchers have not socially coupled this work to an interaction. These various architectures are not designed to take social factors into account such as where to direct the agent's gaze but they have the capability to handle dynamic scenes. The following computational model of directing the robot's gaze through visual attention is inspired by this previous work.

## 6.2 Computational model

This model of attention is based primarily on a competition mechanism that leverages 1) the saliency provided by the signature saliency model, (Hou et al., 2012), 2) previously trained attentional trajectories from a human-environment data collection that provides trajectories that are sensitive to the content in the movie itself and 3) a situated pointing gesture recognition system. While saliency models provide a good representation of where participants may look in a general case, they don't provide a context-sensitive decision on most likely trajectories with respect to time. Using a dataset of participants watching the same video prior to the human-robot interaction experiment allows us to train a most likely estimate fovea trajectory through the environment rather than an estimate of the conspicuousness of various aspects of the scene itself. Previously trained attentional trajectories provide starting locations and transitions that allow the robot to make better decisions about the trajectory of where to look without social intervention. Tsotsos (Tsotsos, 2011) previously described this general design pattern as an architecture in which saliency biases the agent's visual search of the environment. Variance between participants is used to inversely weight the conspicuousness of a specific location at a specific time. When variance is low, saliency

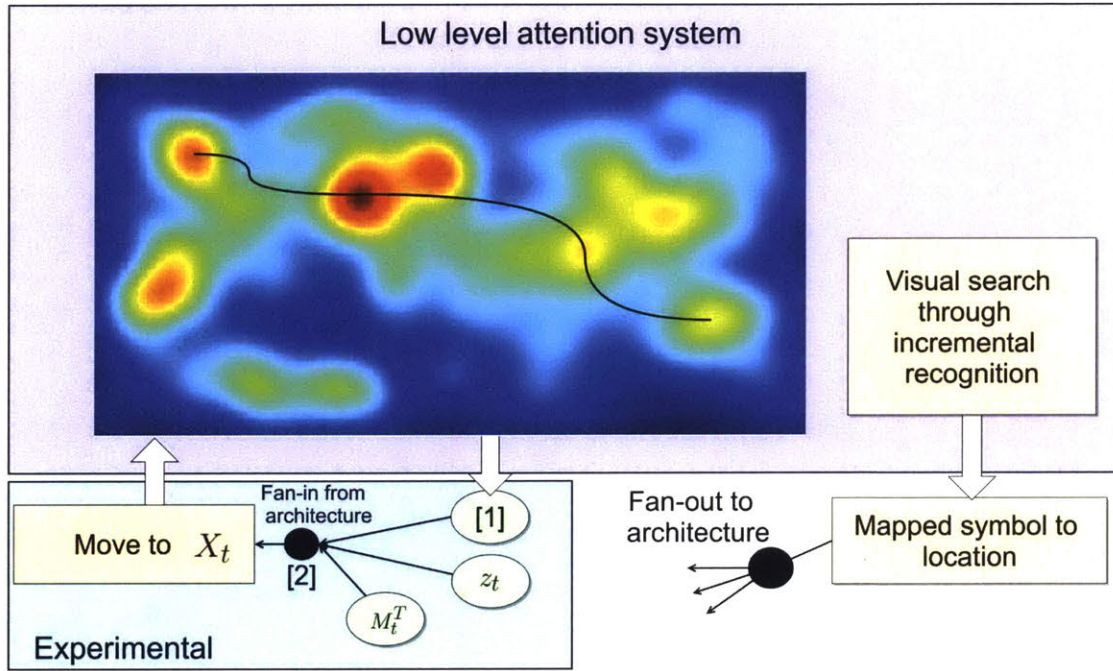


Figure 6-1: Synthetic attention architecture component for interaction. Experimental section marked in blue and the attentional components highlighted in green.  $z_t$  represents the recognized deictic action from the Leap Motion as a binary event ( $z_t \in \mathbb{Z}$ ),  $M_t^T$  represents the trajectory (T) model (M) that returns the location ( $\bar{x}_t = E[x]$ ) as a function of the dataset collected from the NR condition. The [1] function is documented as the NEXTPOSITIONSALIENCY function while the [2] function represents a winner-take-all mechanism (see WTAFOVEA) in Algorithm 6.1.

---

**Algorithm 6.1** Competition resolution algorithm for next foveal position. NEXTPOSITIONSALIENCY returns the local maxima given a random location to choose next. Fixations last for approximately 5s. Typically saliency is weighted lower than the trajectory model but can override the trajectory if variance is high. The weights  $w_s$  and  $w_T$  are tuned manually. The function MAX\_SECOND represents a maximum function that operates on the second element of the list (the competitiveness value). The first function returns the first element of the list (the position).

---

```

1: procedure NEXTPOSITIONSALIENCY(map,  $X_{t-1}$ )
2:                                      $\triangleright$  Next  $X_t$  from saliency
3:    $w \leftarrow \text{width}(\text{map})$ 
4:    $h \leftarrow \text{height}(\text{map})$ 
5:    $x_i \leftarrow \text{rand}(w, h)$ 
6:    $x_t \leftarrow \text{NelderMeadMax}(\text{map}, x_i)$ 
7:   return ( $x_t$ ,  $\text{map}[x_t]$ )
8: end procedure
9:
10: procedure WTAFOVEA( $X_{t-1}$ , sal,  $M_t^T$ ,  $X_t^D$ ,  $z_t$ )
Require:  $\lambda_x \text{sal}_x \leq 1$ 
Require:  $z_t \in \mathbb{Z}_2$ 
11:   if cond = JA and  $z_t = 1$  then
12:      $D_t \leftarrow (X^D, 1.1)$ 
13:   else
14:      $D_t \leftarrow (X^D, 0.0)$ 
15:   end if
16:    $X_t^S \leftarrow \text{NextPositionSaliency}(\text{sal}, X_{t-1}) \cdot [1.0, w_s]$ 
17:   return first(max_second( $X_t^S$ ,  $M_t^T \cdot [1.0, w_T]$ ,  $D_t$ ))
18: end procedure

```

---

may take over, and when interrupt events like deictic gesture locations from human partners arrive, then they override all other factors.

Figure 6.1 describes the system in more detail. A simple saliency model was used to measure the conspicuity of the scene and a feed-forward system that leverages foreground grouping to classify the underlying image. This helps produce meaningful symbols of subfigures and objects based on the classification foreground. More details on this work are found in Chapter 4. Meanwhile, the robot is biased to completely override its attention system when pointing gestures interrupt its gaze in the JA condition (see Section 6.3.1).

The competition was biased for the sake of A/B testing (see Section 6.3). As previously mentioned, the trajectories were trained from previous interactions where time-indexed gaze locations were collected and aligned with the other participants. Next gaze location was interpolated frame-to-frame and their location was averaged for the general case while providing confidence in the form of inverse variance. A probability mass function map was produced using a mixture of gaussians with  $k=2$  and was rendered across the entire scene. This time-indexed saliency map helped make decisions about where to look in winner-take-all (WTA) competition with lower level, normalized saliency maps. Weights were tuned for reasonable behavior.

Joint attention in this architecture was driven by social nonverbal factors as well as goal-oriented visual search. For the sake of the experiment, previously learned gaze trajectories were fed into the architecture to move the fovea. In this work, rather than learning to jointly attend (Triesch et al., 2006; Doniec et al., 2006; Nagai et al., 2003), this system was designed to sustain joint attention in a number of ways: 1) respond to interrupting gesture from the human partner by following gesture, 2) elicit pointing gestures at specific locations at certain times to direct human attention and 3) become distracted from time to time to decouple from the interaction and probe the human participant for a corrective action. The distraction in Point 3 was achieved by overriding the trajectory completely and fixating on a location far off task. This action provided a legibility for the interaction partner who understood the robot had disengaged.

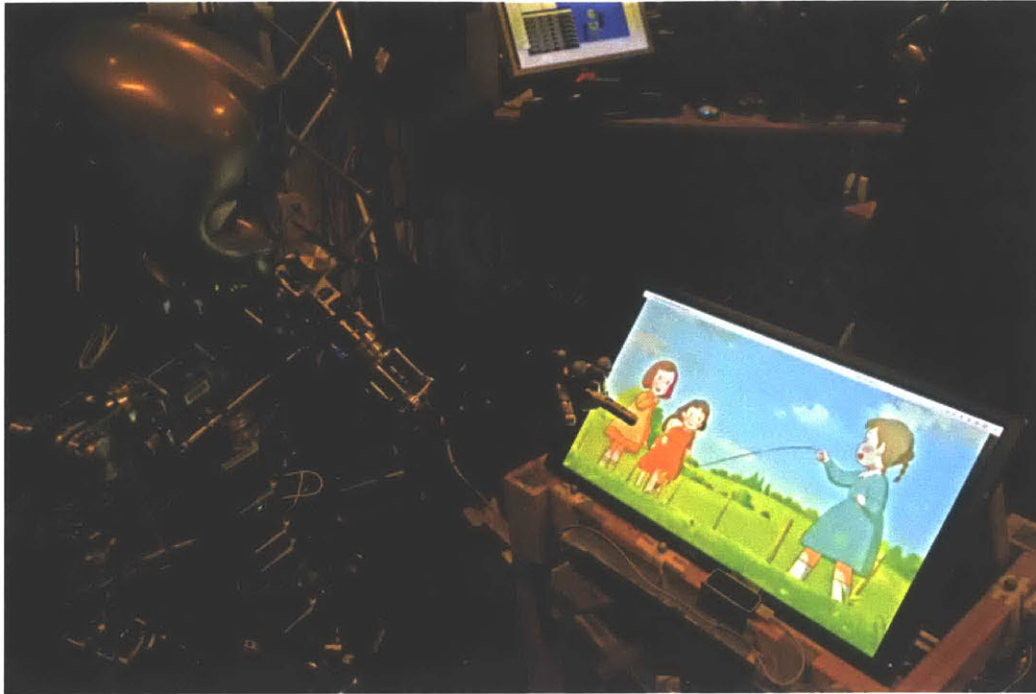


Figure 6-2: Experimental setup with the robot. The robot is an approximately 3.5 foot tall root with fully actuated hands designed to interact with a tilted screen that displays a dynamic narrative. A Leap Motion is mounted just below the display to capture hand gestures. The visual narrative is time synchronized with the robot's deictic action to measure the effect of a particular gesture across the narrative and across the participants. Gaze direction is computed using the method documented in Section 6.2.

### 6.3 Study setup, details and recruitment

To test the hypotheses, a human robot experiment was performed with three chosen conditions. 20 participants per condition were targeted and 72 participants were successfully recruited. Twelve breakdowns occurred which left the 20 target participants. Participants were recruited using local advertisements on Craigslist, email lists targeting local students within the university and advertisements within the department. No (0%) local recruit reported any prior experience interacting with robots (due to the multidisciplinary nature of the department). Additionally, participants recruited from Craigslist reported no experience in interacting with a robot. Participants ages were  $\mu = 39, \sigma = 17$ . The gender breakdown was 63% male and 37% female.

For this experiment, a 57 degree of freedom, approximately 4 foot tall social robot was utilized to interact with the participants (see Figure 6-2). The robot has two arms, each of

which consists of 7 degrees of freedom for the length of the arm and 6 degrees of freedom for each hand. The face has 12 degrees of freedom with cameras mounted in the eyes. The neck consists of 4 degrees of freedom. The robot also has a mobile base that was unused for this experiment.

The software stack consists of a behavior-based animation pipeline used for pointing gestures as well as directing the cameras to specific locations using a tuned head movement mechanism that biases the eye movement prior to head movement. Calibration was performed with simple geometric transformations from a set of Vicon tracked fiducials. Proximal pointing gestures were calibrated using simple animations and hand tuned for precision. To track eye movements, a Tobii eye tracker mounted just below the display was used and calibrated for each participant prior to experiment start. Participants removed glasses on request, only to be contingent on the quality of eye tracking determined by looking at specific points post calibration. Prior to the experiment start, a Leap Motion device tracked hand locations and performed tuning and calibration. Pointing gestures were interpreted as events triggered by the extension of the index finger and calculated as a ray extending from the index finger, originating from the palm center. The Leap Motion provided all points. Assuming the Leap Motion device was parallel to the display, the screen position and normal Leap Motion coordinates were used to calculate the ray-plane intersection showing where the participant was pointing. This point provided a location for the robot to attend to within the screen coordinates.

For the shared content, the experiment utilized Creative Commons material called *Poulette's Chair* (Noitamina, 2014) attributed to "Studio Colorido" (Colorido, 2014) and directed by Ishida YuYasushi. This organization permits the use of the content in educational domains (Google, YouTube. Accessed: 2016-10-03) and allows for modification if the need arises. The content being observed was an animated narrative short featuring a little girl who struggles with socio-emotional issues around making friends, eventually finding a friend in a chair that becomes animate. Using an animated narrative allowed the stimuli being observed to be held constant so analysis of other factors could be performed.

Prior to the experiment, each participant received these simple instructions: "Today you will be watching a movie with our robot, Maddox. Maddox will be watching the movie



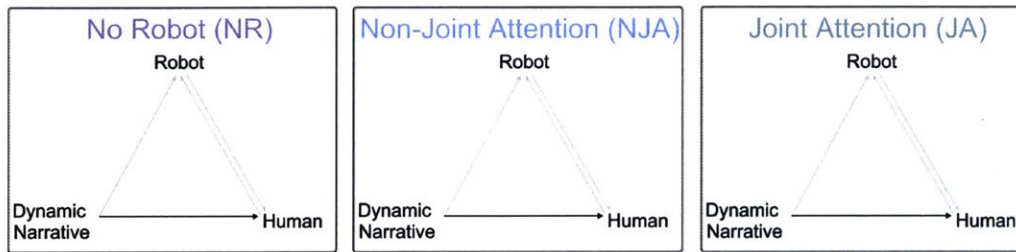


Figure 6-3: Conditions from left to right: No robot (NR) condition, non-joint attention behavior condition (NJA) and joint attention condition (JA). The arrows represent the effect that the factor has on subject. For this experiment, neither the robot nor the human can affect conspicuity outcomes on the content itself.

with you. Your objective is two-fold: you will need to watch the movie and remember as much as possible while also making sure that the robot pays attention. At the end of the experiment, the robot will report about what it saw in narrative form. You, too, will be asked questions about what you saw in the video as well as report as much as you can from what you saw.” The instructions were calibrated based on a pilot to make sure the user 1) prioritized watching the movie, 2) addressed the joint attention problem by understanding that the robot was engaged with the content as well and 3) occasionally might need to direct the robot with referencing behavior.

### 6.3.1 Experimental conditions

Three experimental human-robot conditions were defined corresponding to three basic interaction classes (as in Figure 6-3) that grade the various levels of triadic interactions (environment, human participant and robot). The first condition involved an interaction between an agent (the person) and its environment: the No Robot case (or NR). The second condition focused on incorporating a social partner that was not coupled socially; for instance, the robot did not attempt to adjust the participant’s gaze when he or she wasn’t looking and did not respond to signaling from the participant: the Non-Joint Attention case (NJA). Finally, the last condition was the socially coupled case where the robot responded to the participant’s gesture and elicited gesture toward the directing gaze of the human participant: the Joint Attention case (JA).

Figure 6-3, right, shows the triadic interaction with each condition activating some



subset of the bidirectional process of joint attention. In this case, the environment had an effect on the human subject's gaze patterns. Additionally in the JA condition, the robot may have had an effect on the human's gaze pattern (Hypothesis 1 & 2). In the NJA condition, the robot probed the participant to correct its gaze by looking away briefly. In this case, the expectation was the human would direct the robot back to the shared scene.

*Probes:* To test H5, the robot had a mechanism to probe the user and test whether or not the participant voluntarily remained coupled with the robot. A max time of 18 seconds was set so if a human did not correct the robot's gaze, it would go back to attending to the dynamic narrative. Throughout the experiment, the robot randomly chose a point in time, about  $25 \pm 10$  seconds after the robot attended to the dynamic content. The expectation was the participants would remain coupled, regardless of the condition.

*Directing:* To test H1 and H3, the directing behavior was designed to guide the human participants to a part of the scene that was rarely observed. The NR condition was used prior to the robot conditions to measure gaze trajectories and identify rare events that might be missed by a majority of participants. Next the participants were directed to those locations and asked questions about the content in a post questionnaire. The expectation was if participants saw it, they would remember it. Additionally, they would only see it if they were directed to it, making it more likely that the JA condition would answer the question correctly.

*Post-Questionnaire:* To test H4, a set of questions were posed to the participants, asking them to remember details about the dynamic narrative that the robot directed them toward. They were expected to answer the the questions correctly in the JA condition but not in the NJA condition.

### **6.3.2 Novel quantitative metrics**

To test the hypotheses, two new metrics were chosen to compare saliency and social action. Next the effect of social action was compared. Following are details on how metrics were computed with tests of normality so they could be used in significance tests.

### 6.3.2.1 Measuring the effect of social action on a participant's gaze trajectory: a novel metric

To compare the effect of social action (e.g., pointing on a participant's gaze position in dynamic narrative), a novel metric was designed. The objective of this metric was to specify a way to measure significance between two experimental groups in which two-dimensional gaze position over time is the dependant variable and group is the independent variable. Two groups were compared using a MANOVA test of significance. To generate a summary statistic for a particular participant, the following metric was defined:

$$\bar{x} = \frac{1}{|L|} \sum_{x \in L} x$$

An expected estimate position  $\bar{x} = E[x]$ , is computed across a set of positions  $x$  within a set of frames  $L$ . A set of frames was selected within a specific scene in which the robot might direct its gesture for each participant. Significance was computed as a MANOVA across the expected gaze position within the scene across all participants in a given condition.

### 6.3.2.2 Measuring a participant's saliency: a novel metric

To understand the contribution of saliency to gaze trajectories, the centered saliency that the participant observed was measured. This centered saliency was a measure of currently observed saliency at a specific time. For each gaze position  $x$  within the set of frames  $L$ , the sum of saliency was computed within a small window of size  $w \times h$ .

$$[\sigma_x] = \lambda_{x \in L} \sum_i^{w_x} \sum_j^{h_x} s_{i,j}$$

The resulting set of sum saliency values around  $x$  is measured as  $\sigma_x$  and compared across conditions using non-parametric Mann-Whitney tests.

## 6.4 Results

This section presents the results of testing the hypotheses H1-H5. Each subsection tests one or more hypotheses.

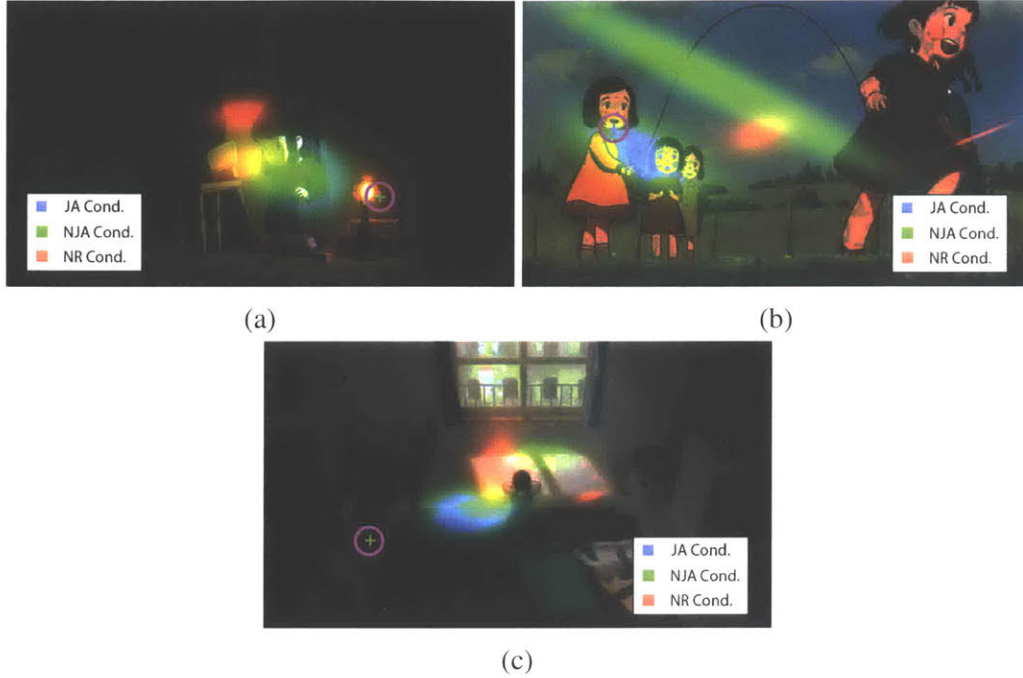


Figure 6-4: Goal target locations the participant was directed toward at various scenes of the visual narrative (annotated with purple). Notice the gaussian in the JA condition is much closer to the directed location compared to the NJA condition.

### 6.4.1 The effect of social action

The first test performed was on the effect of social action: does a robot follow pointing gestures, and if so, how long do people follow gesture? Using the metric defined in Section 6.3.2.1,  $\bar{x}$  is computed. First, each scene where the participant is directed by the robot is measured for average gaze location,  $\bar{x}$ . The central moment of the gaze during these scenes is measured (see Figure 6-4) and is found to be significantly different in every scene in the NJA case (scene 1 (a):  $p \approx 0.03 < 0.05$ ,  $N = 20$ , scene 2 (b):  $p \approx 0.02 < 0.05$ ,  $N = 20$ , scene and 3 (c):  $p \approx 0.043 < 0.05$ ,  $N = 20$ ). Significance is reported from a multivariate analysis of variance (MANOVA) due to the multivariate nature of the dependent variable. Additionally, the differences of whether the participants reported significantly more social behavior from the robot in the NJA vs. JA case are qualitatively reported. Figure 6-5 reports the qualitative difference between the interpretation of the robot's reaction to the participants' commands. This is an important part of sustaining a coupled social interaction with a human participant.

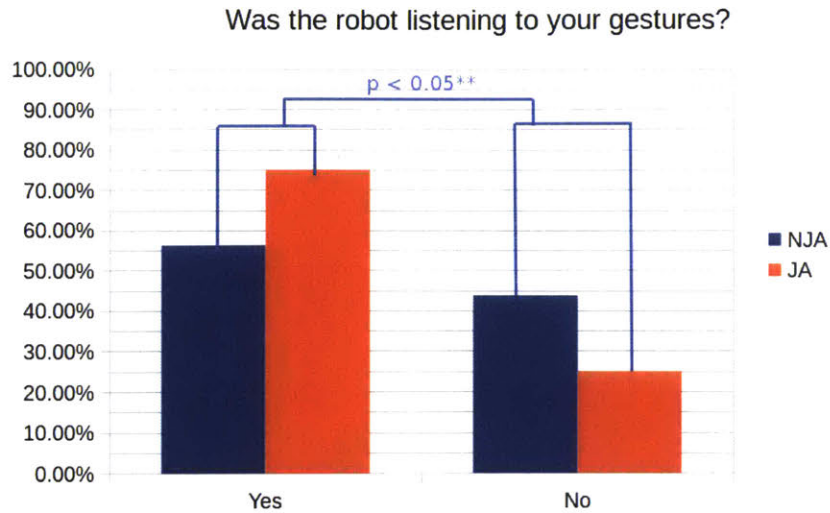


Figure 6-5: A bar chart of qualitative conspicuity of the social (JA) vs. non social (NJA) conditions as reported by the participants. Participants were able to tell a clear difference between the robot that responded to their deictic action and in return, attempted to direct them to the scene at important moments. Significance reported as Student's *t*-test.

Having a dynamic time extended domain also allowed for investigating a more longitudinal set of questions around how sustained JA can be. While robots can theoretically sustain joint attention for as long as possible, there are still questions remaining around how long human participants remain engaged. Yu and Smith regard joint attention as a time-extended exchange of gaze and gesture toward sharing some context (Yu & Smith, 2016). The look-away probe was leveraged to test whether or not the JA coupling had broken. Latency was computed by measuring the time from the onset of a look-away glance to the time of gesture response from the LeapMotion. The accumulated times the robot would probe and no human gesture was detected within the maximum 18s were also computed. Figure 6-6 details both the latency and the misses of the human participant binned over 28s intervals. The findings show two factors be to fascinating: 1) the participant began to respond more tightly as the interaction proceeded, performing better by the end of the experiment than they did at the beginning and 2) the JA condition ended up performing better on average than the NJA condition. This dynamic tuning of their sensory system to respond quickly to cues from the robot is interesting for many reasons. This dynamic tuning to address the situated events suggests the participants adapted quickly and effectively to these low level cues. Findings also show the JA condition outperformed the NJA condition in

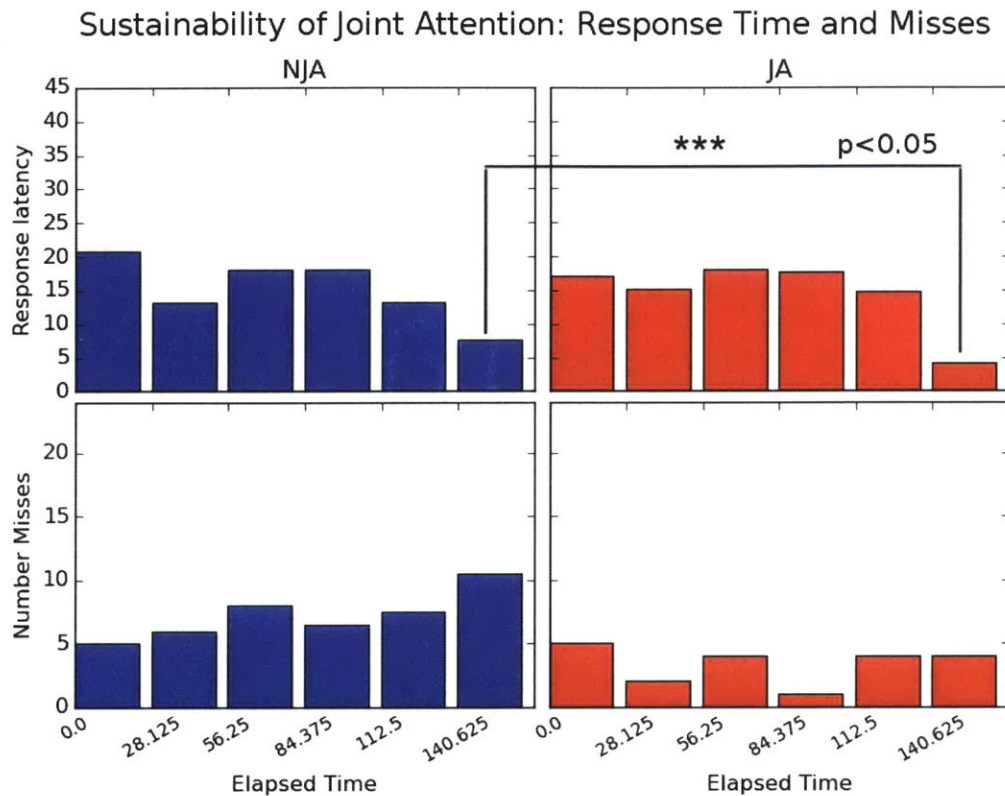


Figure 6-6: Sustainability of Joint Attention: Participants performed reacted in a significantly more timely manner after interaction with the joint interaction robot than with the robot without joint attention. Significance was computed with a Student's *t*-test. Additionally, it is observed that there is a clear trending reaction time improvement over time for both conditions. Finally, participants were more engaged with or chose to engage with the robot more in the joint attention condition. This resulted in fewer missed probes during the joint attention interaction over time and a more sustained interaction. The number of probes missed in the non-joint attention condition increased over time. Significance cannot be reported due to how the misses were computed which resulted in too few samples in both conditions.





Figure 6-7: The participant's mean gaze target (purple) overlaid on top of the saliency map. The purple over the little girl's dress represents the mean gaze target at the moment the robot referenced the auxiliary character on the left. It is clear that saliency of the target region was low at this moment in time compared to the rest of the scene, suggesting that social effects can override the intrinsic conspicuity of the target region.

terms of minimizing the response latency of the social coupling. Finally, though no statistical significance was reported, it seems to be clear and trending that the social coupling in the JA condition also kept the participants engaged by ensuring the participants were monitoring the robot enough not to miss critical probes when the robot became distracted by other stimuli or objectives. The response of the robot to the participant's gesture and the social action taken by the robot suggest that the behavior coupling was important not just for socially inspired robotic cognition but for interaction factors as well.

In the end, the social coupling of the robot's perception system with its human partner had the following effects: a) minimizing the response latency between the participant and the robot and b) coupling the visual search mechanism for both the human participant and the robot. These findings are critical to moving forward with longitudinal interaction.

#### 6.4.2 Effects of social action: a saliency centered analysis

To test hypothesis H2, the observed saliency of each participant group with the tracked gaze position was measured using the Tobii eye tracker against the graph based visual saliency model (Harel et al., 2006) and against the standard Itti-Koch model of saliency (Itti et al., 1998). A simple summary statistic (see Section 6.3.2.2) was computed for each frame of

each participant within a group. The expectation was that the saliency being observed by one group would be substantially smaller than the other group, meaning that robotic social action overrode saliency for brief moments and saliency didn't completely drive the attention of an agent. Indeed, from the three chosen scenes, the null hypothesis was successfully rejected: the accumulated saliency observed in the NJA condition was substantially higher than the saliency observed in the JA condition (MANOVA scene 1:  $p = 0.02$ ,  $N = 2000$  Itti-Koch saliency,  $p < 0.001$ ,  $N = 2000$  GBV saliency, scene 2:  $p < 0.001$ ,  $N = 2000$  Itti-Koch saliency,  $p = 0.008$ ,  $N = 2000$  GBV saliency, scene 3:  $p < 0.001$ ,  $N = 1400$  Itti-Koch saliency,  $p = 0.028$ ,  $N = 1400$  GBV saliency). This suggests that for the human participants, social factors inhibited (through IOR) the saliency of the stimulus enough to allow the deictic gesture to override the desired fixation target. This is an interesting finding for active visual attention systems in which visual understanding drives a substantial set of actions the robot takes (i.e., those designed in (Dickinson, Christensen, Tsotsos, & Olofsson, 1997)) which may be designed to interact with human participants.

### 6.4.3 Memory and performance

While it's been shown in Section 6.4.2 that the robot has successfully directed the gaze location of the human participants, the next question is whether the participants recalled certain details in that location. Hypothesis H4 supposes that once a participant observes a particular location (through some form of JA), they may remember those details. The participants were given a post questionnaire and both the NJA and JA condition were asked a specific question about each scene where attention was manipulated. For scene 1, the question was: what is the color of the dress of the auxiliary character? For scene 2, participants were asked how the little girl lit her room. For scene 3, the question was: where did she keep the chair in her room? All the questions were based on interactions between the participants in the NR condition and the dynamic narrative. The specific details and events were based on rarely observed positions in the narrative.

Figure 6.8 shows that although a plurality of the participants correctly answered every question, there was no significant difference between the two cases. Though evidence

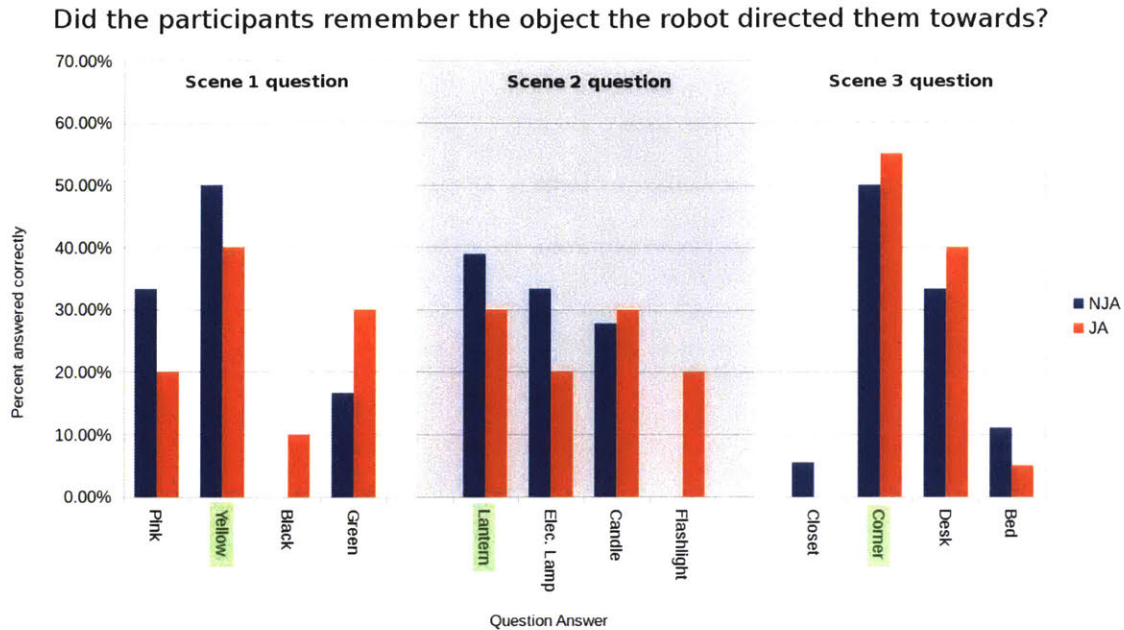


Figure 6-8: Results of human memory recall following study in both the JA and NJA conditions. Unfortunately, no significance was found between the two conditions for any question despite clear evidence that most participants clearly fixated and *saw* the answer.

shows the participants were directed to something in the environment, it's fascinating how they didn't necessarily commit it to memory (called *imprinting*). This has great precedent in basic psychology and this tendency is called *memory failure*. When roboticists give machines perfect or even very good memories, the robots may be able to remember even low level features they observed or *generate them* in a reversible way through generative models. However, it is clear here, more may be needed for the robot to initiate an imprinting into memory.

Following this finding, the research team explored the overall actual performance and potential performance of each participant. The participants in each condition who correctly answered the questions were identified to understand how they might have performed throughout the experiment in other conditions. First, they were separated condition by condition (see Figure 6-9b and Figure 6-9c). While no significance can be reported in this case due to low N, there seemed to be a relationship between latency and misses in those who answered correctly. This question was verified by combining NJA+JA correct vs. NJA+JA incorrect. In this case, a standard t-test was used to verify



## Sustainability of Joint Attention Behavior for High Recall Participants

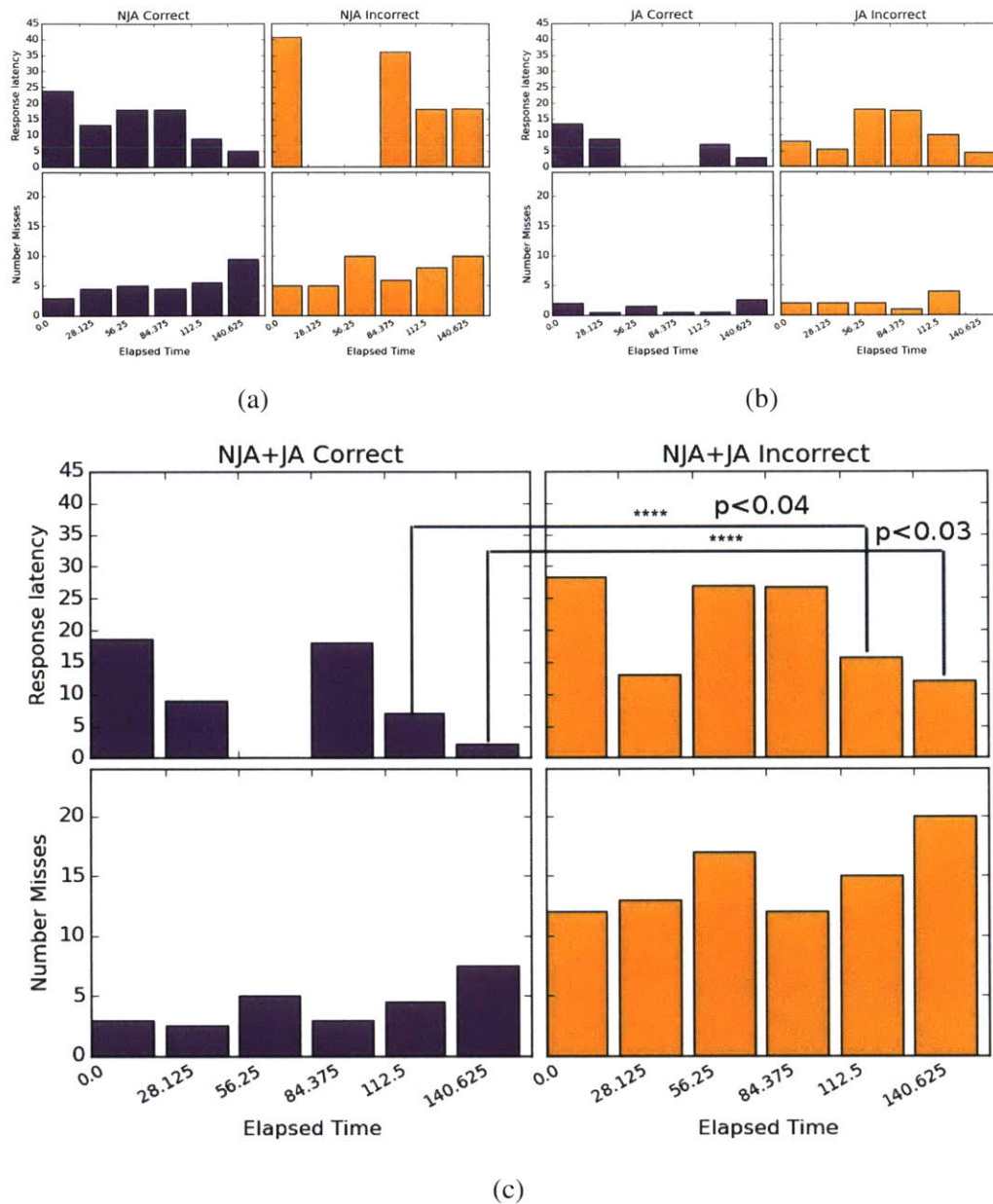


Figure 6-9: An analysis of the performance of the participant's conditional on how they answered the questions in the post-questionnaire. Purple represents those participants who answered the questions at the end of the experiment correctly while orange represents those who answered incorrectly. Top row represents each condition: a) resulting analysis of response times conditional on answering the question correctly. [NJA condition] b) resulting analysis of response times conditional on answering the question correctly. [JA condition] *Note:* Top row's N is too small for significance tests, leading to the aggregate analysis in c. c) the aggregate case of NJA+JA conditional on the participant answering correctly. Notice the tighter coupling from those who answered correctly in the post-questionnaire. Significance was computed using a Student's *t*-test.

( $t \approx (140s, 196s], p = 0.025, N = 17$ ). This finding suggests that those who performed well at coupling the attention of the robot with their own may perform better on other metrics like memory tests.

## 6.5 Discussion

Through the years, deep models of joint attention show up in many areas of interaction design and play critical roles in social cognition. These models have clear interaction effects and one can assume the robot's human counterparts use observation actions to coordinate their motor systems to look at a specific location. However, joint target gazing in educational and collaborative domains should not be the sole phenomenon to rely upon when imprinting details on the human mind. Understanding how to imprint memories on participants is a major agenda item for educational psychology studies worldwide so looking to this body of literature could reveal new theories and a fresh perspective on the problem. One perspective is to view the problem of attention as directly involving cognitive load of the attention system itself (a sort of capacity-centered theory). Barrouillet and colleagues (Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007) have offered a theory which suggests cognitive load can be allocated toward memory as well as other attention demanding processes. In this view, participants may have been under high load during the interaction which has been linked to poor memory commitment (Sweller, 1988, 1994). Meanwhile, specific theories around memory involve list-like recall in which certain memories can decay (J. Brown, 1958) and be overridden by other distractor content during instruction (Fougnie, 2008). These concerns could have also been at play due to the demands of the interaction which may have made memory commitment less likely. Finally Bjork and Whitten (Bjork & Whitten, 1974) have shown that the recency of stimulus can have significant effects on recall. Figure 6.8 demonstrates that the last question asked to the participants about the last scene observed during the interaction had the best recall performance with respect to the other recall tests. This suggests that perhaps the recency of the particular instructional interaction did have an impact on the participants' recall.

While not successful at imprinting memories in the human counterparts, this experi-

ment demonstrates that longitudinal joint attention is sustainable over longer periods of time than previously understood to be possible and it may be critical for long term engagement (a long standing goal of interactive robotics). The experiment found that the more social the robot (the JA condition), the more likely it was for the human participant to monitor the robot to keep the triadic social coupling tight (reducing latency). Additionally, joint attention systems that rely on synthetic visual attention systems to drive the underlying decisions made for the agent should ensure that incoming events such as gestural social action are handled appropriately by reorienting the sensor system to the directed stimulus. It is evident from the JA condition that the attention of the social partner should also be directed when attempting to share the underlying features with the partner. As this study demonstrates, if these problems are not addressed, it seems likely joint attention will break down when the human participant begins to disengage from taking corrective action to direct the robot's sensor system.

### **6.5.1 Unexpected factor: action staging**

One of the most unforeseen and interesting factors to show up in this study is the idea of staging actions. As researchers move into dynamic narratives, timing of actions becomes a premium. To solve this problem, the synthetic attention system had clear attention trajectories through the narrative ahead of time. This researcher was familiar with the content already and also knew to precompute and test it in a controlled experiment. It was understood that the arm and hand must be at a location at a specific time for the robot to direct the gaze location of the participant at the right moment in time. The time it took to extend the arm to reference was measured and the action to execute was tuned prior to directing attention toward a *predicted event*. The action was split into two motion trajectories: one trajectory that extended the arm and positioned it in a *ready-state* and another motion trajectory that extended from *ready-state* into *fully-extended state*. This design is better than the alternatives that make it seem like the robot had the ability to perfectly predict the changing evolution of the environment (the omniscient approach). This strategy allows the robot to stage an action in a ready position for an event it is attempting to predict.

It's interesting that the participants would misunderstand the intention of the action with the 'staged action' design. The robot would move the arm to approximately its waist before extending the arm to action but naive participants frequently misinterpreted this move as "the robot wants to shake hands." As interaction scientists move from static to dynamic joint attention interactions, they should stage gestures ahead of time and communicate the notion that an action is *waiting to execute*. This preparation should ensure participants won't misunderstand these kinematic configurations. In addition, there is also a lack of research on the interpretation of staged actions, an area that warrants deeper investigation and study.

Staged referential action shows up in the general case of dynamic and triadic (environment-peer-self) interactions. To successfully deploy these actions in an interpretable way, roboticians must delve further into the interpretation of these staged actions. For instance, any time environmental prediction is necessary to understand the state of the environment, a decision to reference the participant to a location should be made before the event occurs. Since embodied non-theoretical actions are typically time extended and may be designed piecewise, they also require a deeper reasoning about when to trigger the action. Human interlocutors and partners should carefully consider the interpretation of this type of gesture.

# References

- Admoni, H., Weng, T., Hayes, B., & Scassellati, B. (2016). Robot nonverbal behavior improves task performance in difficult collaborations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 51–58).
- Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3), 570.
- Berlin, M., Breazeal, C., & Chao, C. (2008). Spatial scaffolding cues for interactive robot learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1229–1235).
- Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology*, 6(2), 173–189.
- Bridewell, W., & Bello, P. (2015). Incremental object perception in an attention-driven cognitive architecture. In *Proceedings of the Thirty-Seventh Annual Conference of the Cognitive Science Society* (pp. 279–284).
- Brooks, A. G., & Breazeal, C. (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (pp. 297–304).
- Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10(1), 12–21.
- Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, i–174.
- Colorido, S. (2014). *Chair of Poulette*. <http://colorido.co.jp/works/621/>.

(Accessed: 2016-10-03)

- Dickinson, S. J., Christensen, H. I., Tsotsos, J. K., & Olofsson, G. (1997). Active object recognition integrating attention and viewpoint control. *Computer Vision and Image Understanding*, 67(3), 239–260.
- Doniec, M. W., Sun, G., & Scassellati, B. (2006). Active learning of joint attention. In *6th IEEE-RAS International Conference on Humanoid Robots* (pp. 34–39).
- Dragan, A. D., Lee, K. C., & Srinivasa, S. S. (2013). Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 301–308).
- Fougnie, D. (2008). The relationship between attention and working memory. In *New Research on Short-Term Memory*, ed. Noah B. Johansen, 1, 45.
- Frintrop, S. (2006). *Vocus: A visual attention system for object detection and goal-directed search* (Vol. 3899). Springer.
- Google. (YouTube. Accessed: 2016-10-03). *Youtube help: Creative commons licensing*. <https://support.google.com/youtube/answer/2797468>.
- Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. In *Advances in Neural Information Processing Systems* (pp. 545–552).
- Hou, X., Harel, J., & Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1), 194–201.
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.
- Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., & Breazeal, C. (2013). Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the Conference on Computer Supported Cooperative Work* (pp. 1555–1566).
- Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, 7(2), 135–169.
- Matarić, M. J., & Scassellati, B. (2016). Socially assistive robotics. In *Springer Handbook of Robotics* (pp. 1973–1994). Springer.

- Nagai, Y., Hosoda, K., Morita, A., & Asada, M. (2003). A constructive model for the development of joint attention. *Connection Science*, 15(4), 211–229.
- Nagai, Y., & Rohlfsing, K. J. (2009). Computational analysis of motionese toward scaffolding robot action learning. *IEEE Transactions on Autonomous Mental Development*, 1(1), 44–54.
- Noitamina, S. C. Y. A. (2014). *Noitamina poulette's chair - anime short film hd1080p*. <https://www.youtube.com/watch?v=wg9JGIiSSsQ>. (YouTube. Accessed: 2016-10-03)
- Rolf, M., Hanheide, M., & Rohlfsing, K. J. (2009). Attention via synchrony: Making use of multimodal cues in social learning. *IEEE Transactions on Autonomous Mental Development*, 1(1), 55–67.
- Sauppé, A., & Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 342–349).
- Scassellati, B. (2001). *Foundations for a theory of mind for a humanoid robot* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.
- Seemann, A. (2011). *Joint attention: New developments in psychology, philosophy of mind, and social neuroscience*. MIT Press.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285.
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312.
- Thomaz, A. L., Berlin, M., & Breazeal, C. (2005). An embodied computational model of social referencing. In *IEEE International Conference on Robot and Human Interactive Communication* (pp. 591–598).
- Tomasello, M. (1995). Joint attention as social cognition. *Joint attention: Its origins and role in development*, 103–130.
- Tomasello, M. (2000). *The cultural origins of human cognition*. Harvard University Press.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.

- Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, 9(2), 125–147.
- Tsotsos, J. K. (2011). *A computational perspective on visual attention*. MIT Press.
- Yu, C., Scheutz, M., & Schermerhorn, P. (2010). Investigating Multimodal Real-Time Patterns of Joint Attention in an HRI Word Learning Task. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 309–316).
- Yu, C., & Smith, L. B. (2016). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*.



# Chapter 7

## Reflections and directions

This body of work was specifically conducted to forward researchers' understandings of human-robot joint attention by directly linking human inspired visual attention effects directly with spatial deictic representations. The goal was to couple internal perceptual pointers to objects in the robot's world and use external, directed gestural action toward the robot to direct its perception to the same stimuli. This work defines human-robot joint attention as the ongoing process of social and sequential point-and-recognize loops. This research confirms the use of ambiguous deictic gesture found in previous research (Sauppé & Mutlu, 2014) in which open palm gestures are used in various ways to select multiple targets. This work also contributes a sketch of a visual attention system that attempts to discover and recognize targets, given arbitrary scenes. Using the preliminary evidence, researchers and roboticists can develop a more complex and robust architecture to link visual intelligence systems with deictic exchange in nonverbal communication. This basic architecture could potentially support ongoing verbal exchange by contributing to exophoric reference resolution in a situated manner.

### 7.1 Contributions

The following is a summary of the hypotheses and results from the various studies presented in this document. The hypotheses and findings submitted are constructed as HX.Y where X is the chapter and Y is the hypothesis presented in that chapter.

- H4.1-H4.3: A synthetic attention system with competition at the heart of stimulus mediation predicts the foreground negotiated between a human and a robot. In this case, the mediator was a best fit of the robot taking similar gestural actions to activate the foreground of its partner.
  - Directly selecting the parts of the figure at the point of fixation has many advantages, but one cannot make contextual inferences about connected components of the foreground. On the other hand, prior to referencing behavior, using recognition algorithms helps select the correct referenced figure when the object is already known. However, this approach does not have the ability to cope with novel visual stimuli. Competition, as hypothesized in H4.3, helps mediate the effects to take the best of both worlds.
- H5.1-H5.2: Following deictic action from another agent can be integrated into a machine learning framework so the internal perception system can follow these actions while performing object recognition at the same time. This is specifically demonstrated within a recurrent artificial neural network design.
  - This hypothesis is confirmed and extends work in recurrent attention models that are not trained to respond to social gaze and gesture (Mnih, Heess, & Graves, 2014). Additionally, these models have competitive recognition rates with convolutional neural networks. This model will move a point of fixation toward a referenced location and recognize the object at that location. This extension enhances previous models that are trained to follow gaze to a location without recognition (Triesch, Teuscher, Deák, & Carlson, 2006).
- H6.1-H6.3: The ongoing process of gestural and gaze exchange occurs between a human and a robot in a dynamic environment similar to the triadic interactions observed in relation to static environments.
  - The exchange of gaze and gesture toward dynamic environments resembles the exchange of gaze and gesture in static environments. More specifically, the

gestures must be predicted to map to the environment prior to referencing if the references are to be temporally mapped to the environment appropriately. In this experiment, the solution to the problem involved holding the stimulus constant across all participants with gestures planned ahead of time. There were other interesting gestures that emerged unexpectedly when robots decoupled from the interaction. Namely, participants would trace the spatial line from the robot's center of perception (its eye) to the referent using a finger. It's as though the participant wanted to indicate that the robot should follow the finger itself toward the referent.

- H6.4: The human interaction partner remembered the object the robot gestured toward.
  - I was unable to confirm this hypothesis. Further research is needed to understand why. Section 7.3 discusses some potential avenues for future research in more detail.
- H6.5: The robot's interaction partner will remain engaged throughout the experiment as long as the robot implements the basic interaction components of sharing attention.
  - The exchange of gaze and gesture in a joint attentional coupling may be outwardly related to engagement. In the condition where the robot responded to referential gesturing and attempted to direct the gaze of the participant, gaze cues elicited participant response throughout the interaction. Additionally, in both conditions, participants adapted and improved at responding to the robot. In the case where the robot and participant were tightly coupled, response times and missed reference cues decreased over time consistently.

The findings presented in this document help expand scientists' understanding of the phenomena of joint attention between humans and robots. While further research and engineering is required to implement plausible and working joint attention, it is clear that the behavioral coupling of gaze and gesture between a human and a robot within an environmental reference is critical to ongoing engagement between humans and robots.

## 7.2 Publications

Throughout the work in this dissertation, the following papers and posters were published in collaboration with Cynthia Breazeal:

- Nick DePalma and Cynthia Breazeal. “Object Discovery vs. Selection in Social Action: Benefits of a Competitive Attention System.” In *International Conference on Intelligent Robots and Systems (IROS)*, 2015. (POSTER)
  - This work provided the insights for Chapter 3 and the initial results for Chapter 4.
- Nick DePalma and Cynthia Breazeal. “Sensorimotor account of attention sharing in HRI: Survey and metric.” In *2nd Annual Symposium on Artificial Intelligence and Human-Robot Interaction*, 2015.
  - This workshop paper provided the argument and insights that make up Chapter 3.
- Nick DePalma and Cynthia Breazeal. “Towards bootstrapping socially significant representations through robotic interaction alone: the joint guided search task.” In *Fifth International Symposium on New Frontiers in Human-Robot Interaction at the Artificial Intelligence and Simulation of Behavior*, 2016.
  - This conference paper covered the majority of the results in Chapter 4.
- Nick DePalma and Cynthia Breazeal. “NIMBUS: A Hybrid Cloud-Crowd Realtime Architecture for Visual Learning in Interactive Domains.” In *2nd Workshop on Cognitive Architectures for Social Human-Robot Interaction*, 2016.
  - This workshop paper provided the insights and fundamental system statistics that make up NIMBUS which was documented more fully in Chapter 2.

## 7.3 Recommendations for future work

This dissertation explores two major directions in human-robot joint attention that could lead to a better understanding of the phenomenon in interaction. First, the world is not easily separable into whole parts that are meaningfully aggregated during first contact with new visual stimuli. When parts of objects can be decomposed and recomposed in various fashions, situated nonverbal referencing becomes substantially more ambiguous. In the extreme case, the ability to learn to map reference names for parts leads inevitably to exploring multi-resolution segmentation hypotheses. While this work shows improved understanding in how to measure progress in solving this problem, the door is still left wide open for a solution. The next subsection examines ideas this work did not explore. However, these ideas could potentially lead to solutions and direction on the preference of neural networks to use.

Second, this work provides a study of joint attention in dynamic scenes within the context of learning. I expected many similar challenges identified in earlier experiments of joint attention in static environments but found the dynamic and fast nature of moving scenes to be substantially more challenging for the reasons outlined in Section 6.5. The efficient nature of goal-oriented visual search detailed in Chapter 4 is applicable progress toward a meaningful perception system that could be used in dynamic joint attention domains like those studied in Chapter 6. Joint attention as a phenomenon should be broken down into two component parts for better understanding prior to reassembly: a deeper understanding of the pragmatics (i.e., the observer's understanding) of deictic gesture along with visual attention. Finally, this work makes the case that two cognitive sister fields of attention are critical components for addressing synthetic models of joint attention going forward: artificial context and artificial memory.

### **Shared gaze as a subproblem of visual intelligence**

There is a primary challenge in properly implementing visual attention going forward: investigating the phenomenon requires significant engineering prior to novel research tests. Significant effort will be necessary to understand and build a visual intelligence system

capable of unifying basic recognition and decision making. Chapter 6 showed that participants would direct their gestures toward the sensory center of the robot's perception. This simple observation points to a synthetic ability to handle gestures at a perceptual level. To process these referential gestures, the visual intelligence system will need to recognize a gesture and decide to follow it in meaningful ways. Chapter 5 offers some contributions to this domain in two dimensions; however, it still stands to be extended to handle rays in three dimensions. Finally, recurrent attention models still do not work for objects with three dimensional shape. Let's explore an overview of future work that should be addressed.

1. Chapter 4 and Chapter 5 are meant to be complimentary. To see why, remember that learning in neural networks tend to require templates to be pre-annotated and extracted from the scene with associated labels. Researchers in semantic segmentation have been making steady progress unifying segmentation and recognition for some time, but have not made decisions about where to fixate next. On the other hand, the work of robotics researchers have focused on where to fixate next for some time. The pillars of visual attention must unify before visual intelligence systems will be able to task switch between recognizing gesture, following gesture and recognizing the target.
2. As mentioned previously, recognition and decision making comprise a basic substrate for an artificial visual search mechanism as demonstrated in Chapter 4. To unify this further, models of sequential and potentially hierarchical goal achievement (Erol, Hendler, & Nau, 1994) must be integrated at a higher level of the architecture to direct the fixations and recognition. Additionally, for systems like these to be integrated into contexts that need joint attention systems, further research is needed to unify them into human-robot interaction domains like collaborative robotics. In this researcher's opinion, to address the needs in these domains, interruptible and retaskable mixed initiative planning systems would be a best fit for these attention systems.
3. Chapter 4 explored mediating factors between recognition and learning to unify training and testing within the same framework. Competition in SHARE makes decisions

on when to learn and when to emit recognition signals by deciding if it is sufficiently confident the label being applied by its social partner is toward something already indexed. Competition has been very popular in neural networks through winner-take-all mechanisms like MaxNet, LWTA and max-pooling (Srivastava, Masci, Kazerounian, Gomez, & Schmidhuber, 2013). It is still ambiguous whether these mechanisms can be integrated tightly into a more complex neural architecture. Further research into formalizing the SHARE model into a tightly integrated neural framework could provide a fresh avenue for exploring this work and improving efficiency and accuracy.

4. For this process to work appropriately, memory systems will need to maintain states for long periods of time to ensure proper context switches between tasks. Once gesture is recognized and pointing direction is extracted, memory systems may allow deeper models to cache spatial indices in more meaningful ways for other systems to utilize.
5. There is an abundance of previous work on grounding words to images but many of these systems rely on whole-object biases (Markman, 1990). For robots to move past these assumptions, further research is needed to understand the activation dynamics behind deictic referential behavior. Presently, researchers understand it with the spotlight model, similar to visual attention. Further research in deictic gestures is critical to understanding a more nuanced view of referential gesture. One interesting observation found in Chapter 3 and Chapter 4 was that participants would use piece boundaries as cues to separately reference each part of the whole object. These time extended gestures are meant to be unified into a whole object throughout the demonstration. These effects also need a uniform and simple answer to understand them better.

### **Incorporating visual attention for perception in dynamic scenes**

One of the key advantages of fixation centered tracking and recognition is it represents an efficient solution for real time perception in dynamic environments. The basic idea shows

up in many systems built for tracking and recognition in dynamic environments like VO-CUS (Frintrop, 2006) and the MarVEye system (Dickmanns, 2007). Researchers validated both systems on moving robots in complex moving worlds like autonomous car driving and understanding. As built thus far, the SHARE system can still provide a substrate to make research relevant in these visually guided domains. The research performed to date simulated working systems in these domains and made observations about their effect on interactions. Chapter 6 details how important this ongoing process is for the interaction itself and how without it, social interaction between the human and the robot breaks down quickly. This occurs when human partners begin to understand they are looking but not perceiving the nature of the content at the point of fixation. Two factors are incredibly important in dynamic scenes: 1) tracking movement at the point of fixation and 2) prediction of social (animate) and environmental (inanimate) changes that make action forecasting (the ability to plan an action to have effects at a specific moment) possible.

Finally, high level basic research is required to understand how to perform these tasks in dynamic coupling with the environment.

1. Following up on work in Chapter 4 and Chapter 5, dynamic environments offer a significantly larger challenge for perception systems. Fixation has an even shorter amount of time to emit signals. Following biological inspiration, it's possible tracking must be used as a low-level mechanism to follow stimuli long enough for recognition and understanding to occur. In this case, future work could provide plausible saccade behavior toward the underlying stimulus, allowing the other visual mechanisms like recognition to occur. Following an underlying stimulus would allow higher level mechanisms to work as though they were in more static environments.
2. One of the grand challenges of artificially intelligent systems is in prediction, a new and important domain. There are two key areas in dynamic perception in human-robot interaction: gaze prediction with sparse observation and environmental prediction. When the robot estimates that a human partner is looking elsewhere and the agent must correct its partner to a scene in which objects are moving, it's critical for the system to provide estimates of the other two sides of the triad. For the robot to



discriminate between animate and inanimate objects, it must first recognize them. Following this, the agent must recall which predictive model best previews its next behavior and then update its model. Eye behavior informs much of a social partner's visual experience and new research is shedding light on an approach to using cues to direct eye gaze in marketing and data visualization (Segel & Heer, 2010). This behavior has real cognitive effects on situated understanding, only recently documented in the work of Dr. Neil Cohn (Cohn, 2003, 2013). In addition, this work dovetails with recent research in the neural network community using attention to debug perceptual narrative generation (Xu et al., 2015). Both bodies of work previously mentioned in cognitive and interaction science root the phenomena in narrative understanding: the idea that primitive and predictive shared scripts help fill in gaps in scientists' understanding of an occurrence at any one moment (Bruner, 1991). Scientists and engineers are not entirely certain how to integrate these component parts but new research could help inform this prediction and further illuminate new avenues in mental inference through script inference. More basic research is clearly required before anything conclusive can be said in this space.

## **7.4 Final remarks**

Thanks to research performed in the past two decades, understanding human-robot interaction as a dynamical process between robot and human partner helps inform researchers' hypotheses and drives the interest, engagement and interaction for longer periods of time than previously reported. This work seeks to reproduce some of these mechanisms in meaningful ways and demonstrate the actions of the agent are believable as a synthetic character performing behaviors similar to humans. Conversely, this work strives to not make claims that the synthetic models developed here are the same as those that function in naturally intelligent animals. Clearly though, these phenomena have real effects on both the interaction, behavior and efficiency of an artificially cognitive system. While researchers and roboticists are still a long way away from socially intelligent robots, it is evident that joint attention and joint attention in dynamic scenes will continue to play large roles in under-

standing both nonverbal and verbal linguistic acts by a social partner.

# References

- Bruner, J. (1991). The narrative construction of reality. *Critical Inquiry*, 18(1), 1–21.
- Cohn, N. (2003). *Early writings on visual language*. Emaki Productions.
- Cohn, N. (2013). *The visual language of comics: Introduction to the structure and cognition of sequential images*. A&C Black.
- Dickmanns, E. D. (2007). *Dynamic vision for perception and control of motion*. Springer Science & Business Media.
- Erol, K., Hendler, J. A., & Nau, D. S. (1994). UMCP: A Sound and Complete Procedure for Hierarchical Task-Network Planning. In *AIPS* (Vol. 94, pp. 249–254).
- Frintrop, S. (2006). *Vocus: A visual attention system for object detection and goal-directed search* (Vol. 3899). Springer.
- Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, 14(1), 57–77.
- Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. In *Advances in Neural Information Processing Systems* (pp. 2204–2212).
- Sauppé, A., & Mutlu, B. (2014). Robot deictics: How gesture and context shape referential communication. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (pp. 342–349).
- Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 1139–1148.
- Srivastava, R. K., Masci, J., Kazeroonian, S., Gomez, F., & Schmidhuber, J. (2013). Compete to compute. In *Advances in Neural Information Processing Systems* (pp. 2310–2318).
- Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not)

learn it? *Developmental Science*, 9(2), 125–147.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., . . . Bengio, Y. (2015).

Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3), 5.

# **Appendix A**

## **Datasets used throughout the dissertation**

The following datasets were used throughout the dissertation. Figures A-1, A-2 and A-3 is a small capture of the dataset taken on October 15th, 2015. This dataset was used throughout Chapters 4 and 5.

## A.1 The tangram dataset (snapshot October 15, 2015)

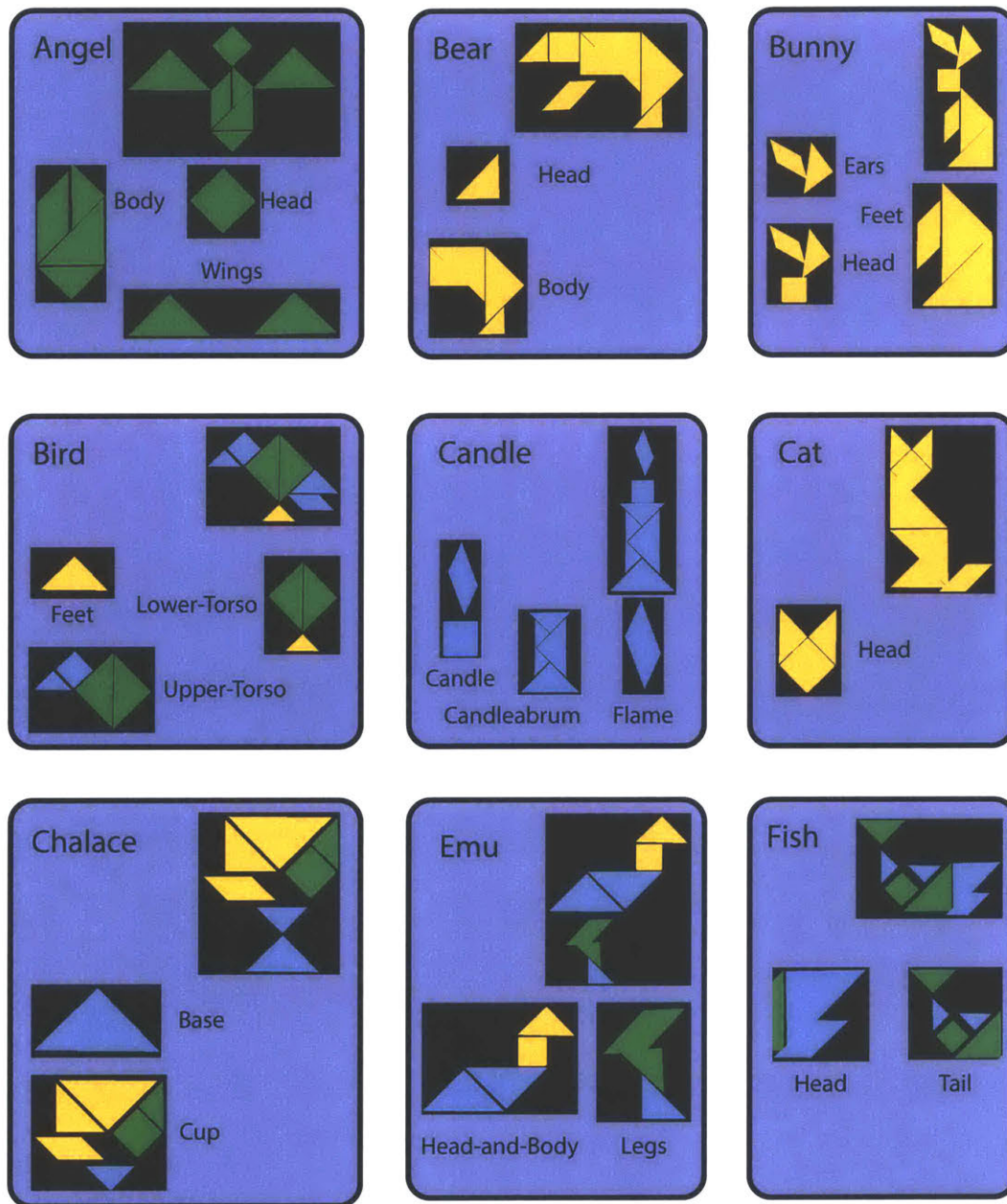


Figure A-1: Sample 1 of the tangram dataset captured from participants on-line. Both whole figures are shown here as well as the subfigures that were selected and labeled by the same participants.

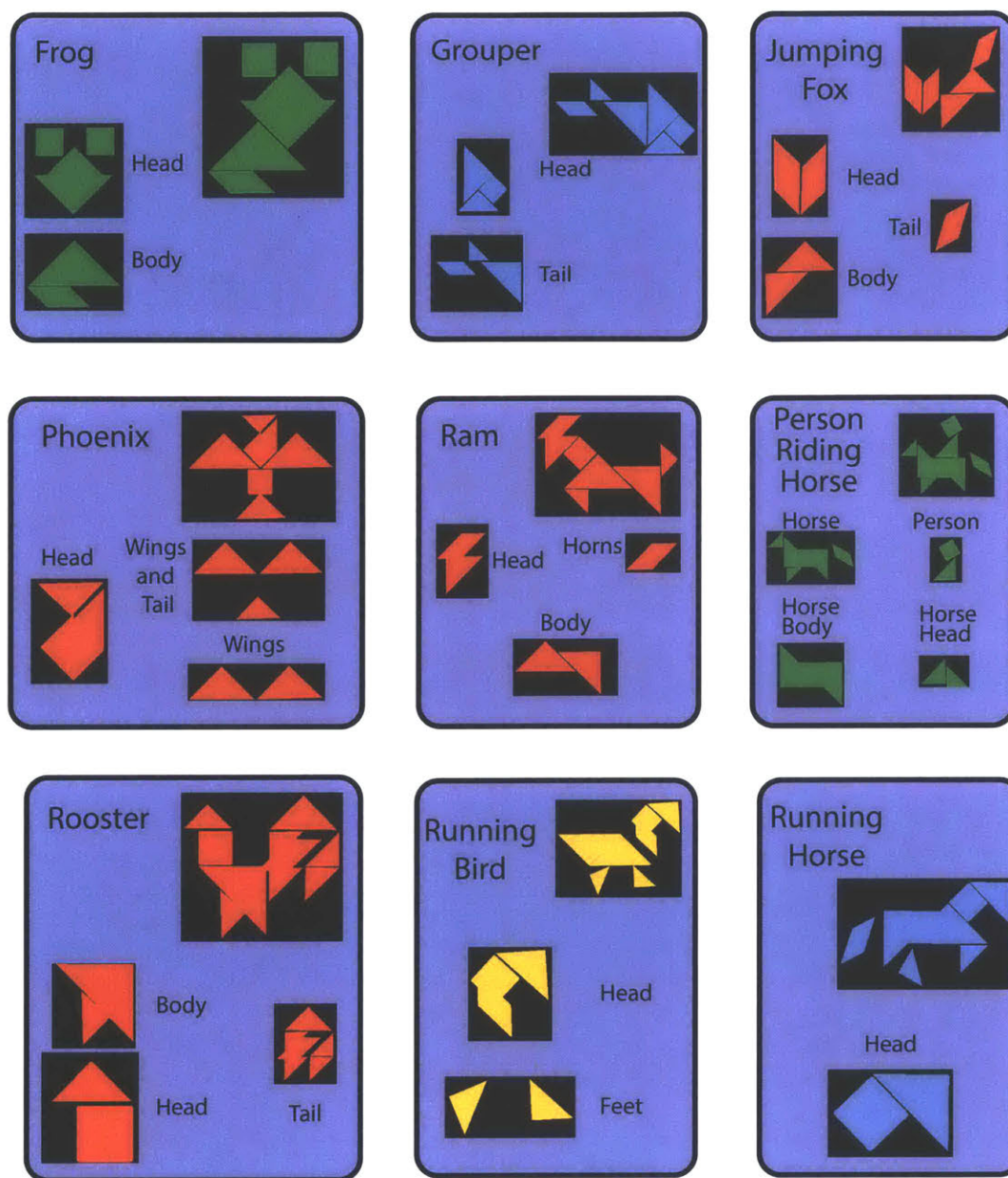


Figure A-2: Sample 2 of the tangram dataset captured from participants on-line. Both whole figures are shown here as well as the subfigures that were selected and labeled by the same participants.



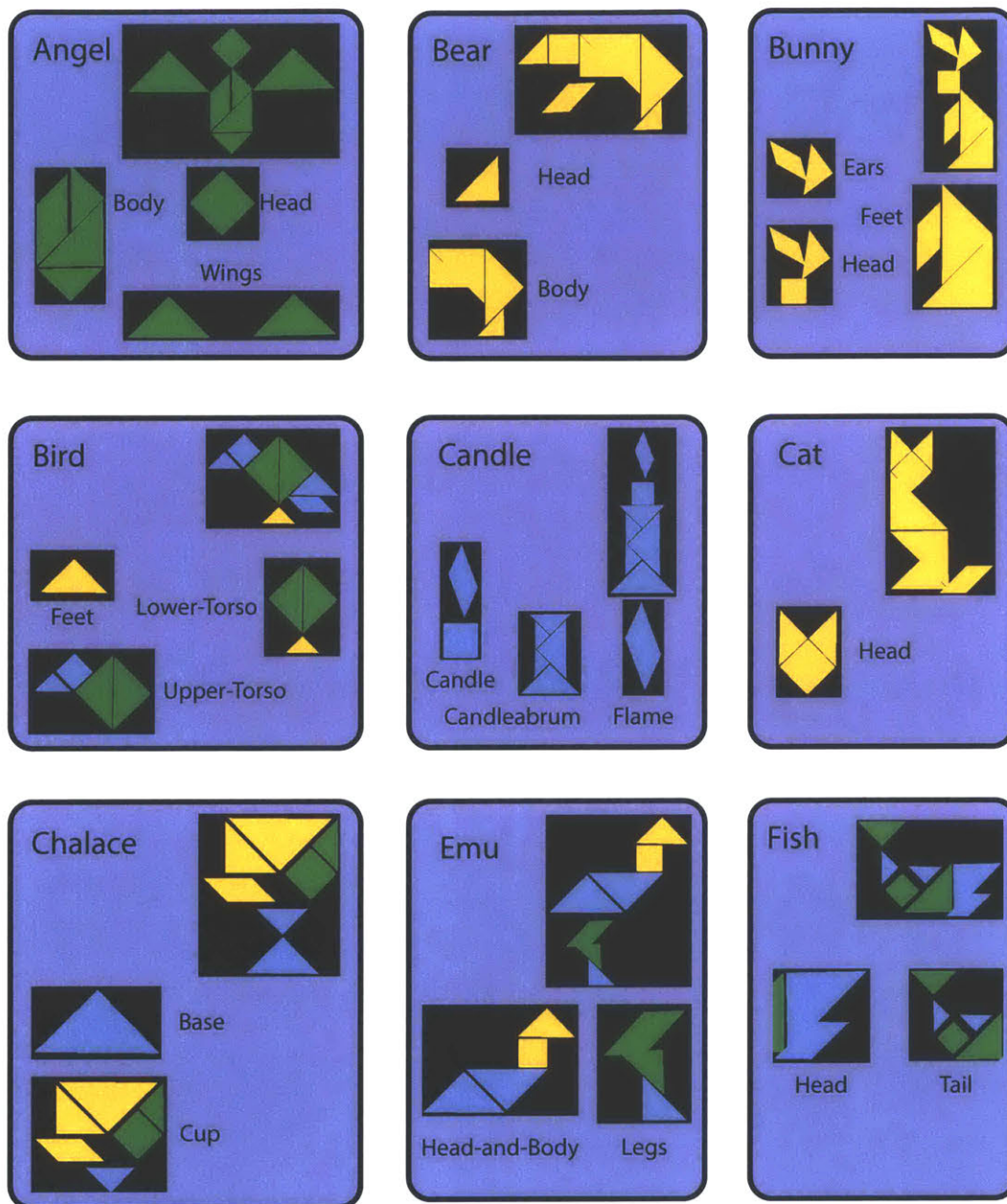


Figure A-3: Sample 3 of the tangram dataset captured from participants on-line. Both whole figures are shown here as well as the subfigures that were selected and labeled by the same participants.



## **Appendix B**

### **Recruitment documentation and consent forms**

The following consent forms were signed and agreed to by all participants interacting with the SHARE and NIMBUS system. Instructions were stated as printed to participants prior to the interaction and a short question period was held to allow for clarifications. The questionnaire documented in Section B.3 was administered to participants who interacted with the robot, Maddox for the final study discussed in Chapter 6.

## B.1 Consent forms

3/10/2017

Consent Page

o

### Consent Page

Posted on Sep 27, 2015

#### CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH

##### Software Game Character Training Experiments

You are asked to participate in a research study conducted by Cynthia Breazeal (Associate Professor), Jin Joo Lee (Ph.D. candidate) and Nick DePalma (Ph.D. candidate) at the Massachusetts Institute of Technology (M.I.T.). Results of this study will contribute to the Ph.D. thesis research of Jin Joo Lee and/or Nick DePalma. You were selected as a possible participant in this study because you have volunteered to label data online with our system and you are over the age of 18 and are eligible to consent. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

#### PARTICIPATION AND WITHDRAWAL

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so. This may happen if foul language is used or if it is found that the participant is undermining the study through intentional deceit.

#### PURPOSE OF THE STUDY

We are investigating Machine Learning applications for software computer games.

#### PROCEDURES

If you volunteer to participate in this study, we would ask you to do the following things:

You will be asked to interact with a tangram game or scene in which a virtual robotic agent is jointly interacting. You will be able to communicate with the character through the use of the keyboard and the mouse. You will be told when the study ends and will then be asked a series of questions about your experience. In some cases, you will have the opportunity to voluntarily exit at any time but will still be required to fill out the questionnaire. The complete study is estimated to take less than one hour of your time.

#### POTENTIAL RISKS AND DISCOMFORTS

We are unaware of any potential risks in this experiment. To reiterate, you may exit at any time you feel discomfort and we would like to request that you report any discomfort immediately to one of the investigators through our contact information below.

#### POTENTIAL BENEFITS

file:///home/nd/projects/turkjs/public/story-collect/\_site/consent/2015/09/27/alpha.html

1/3

Figure B-1: Page 1 of the consent form used throughout the experiments discussed in Chapters 2 and 4. This consent form was covered under COUHES protocol number 1402006168. The protocol was titled “Using Crowd-Sourcing to Understand Effects of Narrative on Perception in Human-Robot Interaction.”

Your participation will help us to build software agents and robots that are more responsive and sociable learning partners.

## **PAYMENT FOR PARTICIPATION**

Every participant will receive, \$0.10 for doing the experiment. You can receive up to an additional \$1 based on the quality of your response.

## **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

While we do not record any identifying data from you, the participant, any information that is obtained in connection with this study will remain confidential and will be disclosed only with your permission or as required by law.

Every participant's session with this online system is anonymized using globally unique identifiers in conjunction with unique identifiers that relate to the study itself. We never will ask for your name, address, or any identifying information. Data is stored in a database that we use to train our robots and will only be used to generate robotic behavior. Any other uses of the data will require further consent of the participant.

## **IDENTIFICATION OF INVESTIGATORS**

If you have any questions or concerns about the research, please feel free to contact:

Nick DePalma (Ph.D. candidate)  
617 452 5603  
MIT Media Lab  
E15-485  
Cambridge, MA 02139  
ndepalma@media.mit.edu

Associate Professor, Cynthia Breazeal  
617 452 5601  
MIT Media Lab  
E15-468  
Cambridge, MA 02139  
cynthiab@media.mit.edu

## **EMERGENCY CARE AND COMPENSATION FOR INJURY**

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

Figure B-2: Page 2 of the consent form used throughout the experiments discussed in Chapters 2 and 4. This consent form was covered under COUHES protocol number 1402006168. The protocol was titled "Using Crowd-Sourcing to Understand Effects of Narrative on Perception in Human-Robot Interaction."

## RIGHTS OF RESEARCH SUBJECTS

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

## RESEARCH SUBJECT AGREEMENT

I understand the procedures described above. By clicking on the *Agree* Button below, you are agreeing that you are of legal age to consent (18 years or older) and that you agree to the above terms.

[Agree](#) [Leave Immediately](#)

Theme Copyright © [Camille Diez](#) 2015  
Powered by [Jekyll](#)  
Subsystem by [Nick DePalma](#)

Figure B-3: Page 3 of the consent form used throughout the experiments discussed in Chapters 2 and 4. This consent form was covered under COUHES protocol number 1402006168. The protocol was titled “Using Crowd-Sourcing to Understand Effects of Narrative on Perception in Human-Robot Interaction.”



## Epsilon Page

Posted on Sep 27, 2015

**Thank you for participating in our experiment.**



Theme Copyright © [Camille Diez](#) 2015

Powered by [Jekyll](#)

Subsystem by [Nick DePalma](#)

Figure B-4: The webpage presented to participants on-line who have already participated in the study.

## **CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH**

### Dynamics of Social Interaction and Cooperation between Humans and Robots

You are asked to participate in a research study conducted by Dr. Cynthia Breazeal, and Nick DePalma from the Program in Media Arts and Sciences at the Massachusetts Institute of Technology (M.I.T). You were selected as a possible participant in this study because you are at least of 18 years of age and have not been diagnosed with an autistic spectrum disorder. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so.

- **PURPOSE OF THE STUDY**

The purpose of this study is to investigate the interaction between humans and between a robot and a human. We are interested in how people communicate and interact with each other and with robots, and we hope that the results of this study will help us to improve the design of the robot's collaborative and social abilities.

- **PROCEDURES**

If you volunteer to participate in this study, we would ask you to do the following things:

You will be asked to participate in a one-on-one social interaction with either another participant or with a humanoid robot. The participant will be asked to engage in a 10 to 30 minute interaction with the robot or with another participant. You will be instructed what types of nonverbal gestures you can use. If you are interacting with a robot, understand that it does not understand your words and will not speak back. If you are interacting with another participant, no talking will be allowed.

You will be asked to complete a questionnaire after the session. After you complete your questionnaire, you can ask questions about the experiment, the robot, and the goals of the research.

- **POTENTIAL RISKS AND DISCOMFORTS**

There is a minimal risk that you might feel uncomfortable speaking to the other participant or to the robot. If any part of the conversation is in any way uncomfortable, you can opt to change the topic or end the interaction.

Figure B-5: Page 1 of the consent form used throughout the experiments discussed in Chapters 4 and 6. This consent form was covered under COUHES protocol number 1403006294. The protocol was titled “Dynamics of Social Attention and Cooperation with Humans”.

- **POTENTIAL BENEFITS**

There are no specific benefits that you should expect from participating in this study; however, we hope that you will find the experience to be enjoyable and engaging.

Your participation in this study will help us to build robots that are better able to interact with humans. And the information learned from this study may advance scientific understanding of interpersonal interaction

- **PAYMENT FOR PARTICIPATION**

You will receive \$10 at the conclusion of the experiment and a chance to win \$30 gift certificate to a vendor of your choice. There is no penalty for withdrawing before the end of the study.

- **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

Your participation in this study will be record by regular cameras and by 3D cameras. The data will be kept in a locked, secure location after the conclusion of the study. No data that would describe an individual participant will be used, we will only use aggregate data from all participants.

At any time during or after the experiment, you can request that all data collected during your participation be destroyed.

- **IDENTIFICATION OF INVESTIGATORS**

Associate Professor Cynthia Breazeal  
617-452-5601  
MIT Media Lab  
E15-468  
Cambridge, MA 02139  
[cynthiab@media.mit.edu](mailto:cynthiab@media.mit.edu)

Nick DePalma  
617-452-5112  
MIT Media Lab  
E15-468  
Cambridge, MA 02139  
[ndepalma@mit.edu](mailto:ndepalma@mit.edu)

- **EMERGENCY CARE AND COMPENSATION FOR INJURY**

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

Figure B-6: Page 2 of the consent form used throughout the experiments discussed in Chapters 4 and 6. This consent form was covered under COUHES protocol number 1403006294. The protocol was titled “Dynamics of Social Attention and Cooperation with Humans”.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

- **RIGHTS OF RESEARCH SUBJECTS**

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787

Figure B-7: Page 3 of the consent form used throughout the experiments discussed in Chapters 4 and 6. This consent form was covered under COUHES protocol number 1403006294. The protocol was titled “Dynamics of Social Attention and Cooperation with Humans”.



## B.2 Instructions provided to participants

In this study, you will be helping us understand how you select components of a scene. To help us out, one of you must show the other what he/she is looking at. That person will have the true answer on their screen.

The instructions are as follows:

1. I will place a piece of paper between you both. One person will need to communicate to the other which parts of the tangram scene they must select.
2. Once the paper comes down, one person will receive the scene as is and the other with the correct parts selected.
3. You will have 5 seconds to communicate the selection.
4. Once those 5 seconds expire, the receiver of the selection will need to enter in the selection on the screen.

Further rules of the game:

1. There is no talking.
2. You are both allowed to gesture if you would like.
3. You will have 5 seconds only beginning on the first gesture exchanged.
4. Part 2 of this study will be performed on both of you individually.
5. You are allowed to touch the paper but not rotate it or move it.
6. On the displays
  1. highlighted or *selected* tangram shapes are outlined in white on your screen
  2. tangram shapes not highlighted or selected are outlined in black on your screen

For the receiver of the experiment, you should be able to enter your selection by pressing the display with your finger.

Once you are finished with a scene, press the next button to advance the study.

This will repeat until the administrator indicates the end of the study.

Figure B-8: The instructions provided to the participant dyad prior to exchanging gesture for the study discussed in Chapter 3.

In this study, you will be *sharing attention* between the robot and yourself. *You will receive a selection that you must communicate* that is framed as a “foreground” or “selection” of the object you are looking at. On the display to your left, two windows will be open. The *window on the left shows you the foreground you should communicate* to the robot. The window on the left will allow you to *enter the selection you think the robot is attempting to communicate back*. You may not verbally communicate with the robot and *you may only communicate with the robot through nonverbal, referential, gesture* in the holographic illusion between the robot and yourself.

There will be **three phases**.

Phase 1: The robot is passive, **you will gesture toward the robot to communicate the goal**. The robot will perceive your gesture and make a guess at your selection.

Phase 2: The robot will gesture back to you to communicate the goal. **You must then enter the foreground on the touchscreen to your left.**

Phase 3: You will both receive a foreground to communicate. **You and the robot will gesture to each other to communicate the foreground**. You will report both what you think the robot is communicating as well as what you think the shared foreground is.

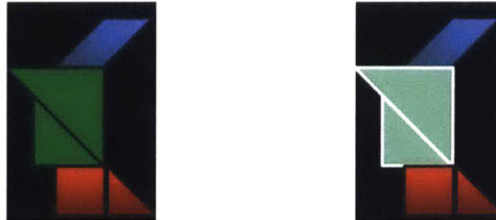


Figure 1: (Left) The scene presented to you and the robot. (Right) The foreground selected that you must communicate to the robot.

Figure B-9: The instructions provided to the participants interacting with Maddox for the study discussed in Chapter 4.

## B.3 Questionnaires

3/10/2017

Chair Survey N

1. What happened in the video?

2. What story do you think the robot would tell based on your interactions with the robot?

3. What is your identifier?

4. Please rate your experience with the interaction on a scale of 1 to 7.

	Very frustrating			Neutral			Pleasant experience
Did the robot follow directions?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do you think that I was controlling the robot?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do you think the robot was trying to direct you?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5. Do you believe that the robot was listening to you? (i.e. following your gestures or gaze?)

<https://www.surveymonkey.com/r/TWM5KZ5>

1/1

Figure B-10: Page 1 of the questionnaire administered to the participants interacting with Maddox following the study discussed in Chapter 6.

3. What is your identifier?

4. Please rate your experience with the interaction on a scale of 1 to 7.

	Very frustrating			Neutral			Pleasant experience
Did the robot follow directions?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do you think that I was controlling the robot?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Do you think the robot was trying to direct you?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

5. Do you believe that the robot was listening to you? (i.e. following your gestures or gaze?)

☐ Yes

☐ No

Other (please specify)

6. Is there anything you'd like to add about how the robot responded to your gesture?

Done

Figure B-11: Page 2 of the questionnaire administered to the participants interacting with Maddox following the study discussed in Chapter 6.