Modeling the Dynamics of Nonverbal Behavior on Interpersonal Trust for Human-Robot Interactions

by

Jin Joo Lee

BSc. Electrical Engineering, Georgia Institute of Technology (2008)

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

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August 2011

Author_____

Program in Media Arts and Sciences
August 5, 2011

Certified by_____

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

Accepted by_____

Prof. Mitchel Resnick
Academic Head, LEGO Papert Professor of Learning Research
Program in Media Arts and Sciences

Modeling the Dynamics of Nonverbal Behavior on Interpersonal Trust for Human-Robot Interactions

by

Jin Joo Lee

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 5, 2011, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

We describe the design, implementation, and validation of a computational model for recognizing interpersonal trust in social interactions. We begin by leverage pre-existing datasets to understand the relationship between synchronous movement, mimicry, and gestural cues with trust. We found that although synchronous movement was not predictive of trust, synchronous movement is positively correlated with mimicry. That is, people who mimicked each other more frequently also move more synchronously in time together. And revealing the versatile nature of unconscious mimicry, we found mimicry to be predictive of liking between participants instead of trust. We reconfirmed that the following four negative gestural cues, leaning-backward, face-touching, hand-touching, and crossing-arms, when taken together are predictive of lower levels of trust, while the following three positive gestural cues, leaning-forward, having arms-in-lap, and open-arms, were predictive of higher levels of trust. We train and validate a probabilistic graphical model using natural social interaction data from 74 participants. And by observing how these seven important gestures unfold throughout the social interaction, our Trust Hidden Markov Model is able to predict with 94% accuracy whether an individual is willing to behave cooperatively or uncooperatively with their novel partner. And by simulating the resulting model, we found that not only does the frequency in the emission of the predictive gestures matter as well, but also the sequence in which we emit negative to positive cues matter. We attempt to automate this recognition process by detecting those trust-related behaviors through 3D motion capture technology and gesture recognition algorithms. And finally, we test how accurately our entire system, with low-level gesture recognition for high-level trust recognition, can predict whether an individual finds another to be trustworthy or untrustworthy.

Thesis Supervisor: Dr. Cynthia Breazeal

Title: Associate Professor of Media Arts and Sciences, Program in Media Arts and Sciences

Modeling the Dynamics of Nonverbal Behavior on Interpersonal Trust for Human-Robot Interactions

by

Jin Joo Lee

The following people served as readers for this thesis:

Thesis Reader_____

Dr. David DeSteno
Associate Professor of Psychology
Northeastern University

Thesis Reader_____

Dr. Rosalind Picard
Professor of Media Arts and Sciences
Program in Media Arts and Sciences

Acknowledgments

First and foremost, I want to thank my advisor Cynthia Breazeal for all the encouragement and support she has given me for the last 2 years. She has pushed me and challenged me into becoming a better researcher and has inspired me with her ability to drive our group towards visionary directions. She is truly a great role model, and with her guidance I am one step closer in becoming a world-class roboticist.

David DeSteno has been a wonderful mentor and a great advisor in introducing me to this fascinating world of human social psychology. It has been a privilege working with him and his students Jolie Baumann and Leah Dickens - an experience that has taught me the true meaning of collaborative work.

I am very grateful in having an extraordinary professor like Rosalind Picard be a part of my research career. Even in her busiest of schedules, she always managed to find time to give invaluable feedback on my work. And without the machine-learning material she has taught me, I would not have been able to accomplish some of the critical work in this thesis. I also want to thank her students Dan, Ehsan, and Rob for helping me with the numerous questions I have asked for the past 2 years.

Although my current advisors are the ones that drive me forward today, I cannot forget the team of advisors at Georgia Tech that have shaped my collegiate life into being prepared for my graduate career at MIT. Thank you Dr. Ayanna Howard for being my very first robot advisor that sparked my robot interests, and Sekou Remy for encouraging me to continue and persist in doing research. Thank you Dr. Andrea Thomaz for kindling this spark into a flame with our worked with Simon the robot. Dr. Mimi Philobos for encouraging me and giving me, a woman in engineering, endless support to excel in a male-dominated field. And lastly, Dr. Patricio Vela for introducing me to how much fun hands-on robotics can be and for having such sincere interest and concern for my academic success and future plans. Thank you for spending so much time explaining complicated concepts to me that I did not get the first time, or the second.

The staff at the Media Lab has also been an essential support for me in the last 2 years. True student-advocates, Linda Peterson and Aaron Solle were always so helpful, and they both really look out for the well-being and sanity of the graduating class. Polly Guggenheim gave endless supporting hugs and encouraging words, and she always knew the right thing to do even at the most despairing of times. She is the heart that keeps the robot family alive and pumping.

I have had the extreme privilege in calling the Personal Robots group my home for the past 2 years, and by working and learning from my fellow colleagues, I have grown into a better roboticist. Thank you Nick for all the endless coffee conversations where we dream and fantasize about the robot future and then strategize our way towards it. You have not been only a great colleague to me since SIM lab but also a great friend. Julián without all your help with the vision framework I would not have been able to get my study up and running on time. Thank you for spending

so much time and effort in helping me code and debug our work. Alumni Jesse and Matt, you guys really set the collaborative tone and the high bar of excellence for the robot group. Thank you guys so much for helping me call both robots and Boston my new home. Sonia, not only are you a dear friend but you are also such an inspiration to me as fellow female roboticist. Even when you are so busy and so far away, you always made time to have lunch with me to just talk about life. Jason, thank you for checking up on me to make sure I still had my sanity and making reasonable progress on my thesis. I always enjoy our talks about the future of artificial intelligence over 3pm clover snacks. David and Peter, you both have been the best kind of officemates anyone could ask for. We all suffered together through our thesis and came out a lot closer to each other. And Peter thank you so much for sitting through and patiently listening to all my endless complaints and life drama. Siggi, thank you for being such a good role model in showing what teamwork, leadership, and being a badass roboticist is all about. Adam S. and Natalie, you guys really embraced and exemplified what it means to have the robot spirit. I am so glad and honored to have you guys a part of the robot family. And thank you so much Angela, Kenton, Kris, Nancy, Phillip, and Adam W. for always being so eager to lend a helping hand or an open ear. I love my robot family.

Lastly and most importantly, my family and friends thank you so much for all your emotional support and happy memories. Sunny, you have been such a dear friend to me since the 6th grade. We managed to travel through and survive academia together, and I am very lucky to have such a kind and caring childhood friend like you. Santiago and Caitlin, you guys always remind me to live life to the fullest and to always have fun no matter what, even if that requires me to abandon my desk. Cranor, Edwina, Xiao Xiao, and Carr, we made it through the masters program together, and I am so glad I had you all through this process. And Jorge, thank you for pushing me to become a better person everyday and for always believing that I could pull through out of any situation.

Jin, even with all our sibling rivalry, we managed to become best friends, and we both have become very successful women in our careers. Thank you so much for being such an awesome older sister for the last 24 years of my life. Mom and Dad, I know you are both so proud of us, and we thank you for all the sacrifices you have made. Your hard work is what really inspired us to relentlessly climb towards our dreams.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

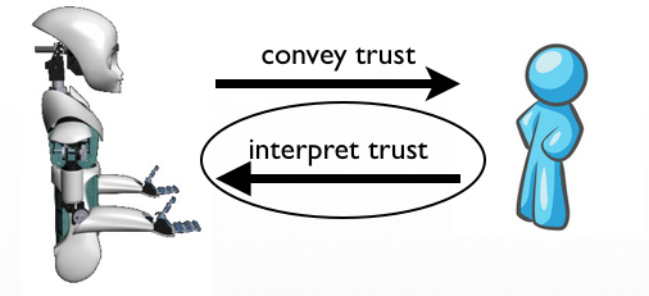
Honest signals are primitive social signals communicated between people through unconscious behaviors and are an effective window into our intentions, goals, and values [Pentland, 2008]. By observing these unconscious behaviors, we can gain insight into how well a date is going [Pentland, 2008], how successful of a deal will be made [Maddux et al., 2007], and the impressions of an interviewer [Soman, 2009]. So much of our communication is beyond words that by ignoring nonverbal behavior we create socially ignorant technology is completely frustrating; and robots are of no exception.

Robots have an immense potential to help people in domains such as healthcare and education, and their success heavily depends on their ability to effectively communicate and interact with us. As robots begin to closely work with us, we need to consider some very important mediating factors that can affect the outcome of the human-robot interaction. Interpersonal variables like trust, friendliness, engagement, rapport, and comfort need to be designed in such a way that is appropriate in different contexts.

In a context like healthcare, where robots are being used to collect personal and sensitive information from patients, trust, in particular, is an essential variable to consider [Wilkes et al., 2010]. Studies have shown that increased levels of trust facilitates open communication and leads to more information sharing [Maddux et al., 2007]. Thus, when gathering personal medical information, robots will need to establish a sense of trust with the patients to allow for more effective communication and exchange of information.

And in forming even closer relationships, robots are starting to be used as potential social partners and peer-tutors for children. In the context of education, where interactive robots are being used as second-language teachers in elementary schools

[Kanda et al., 2004], trust again is an important factor to consider as the student-teacher trust relationship is a key predictor of a child's academic motivation, performance, and school adjustment [Lee, 2007].



When designing for such interactions, we need to answer how can a robot:

1) **convey** the appropriate level of trust to an individual (the signal)
2) **interpret** how much an individual trusts the robot (the feedback)

The cumulation of work presented in this thesis is an effort in answering this second question of how a robot can understand, or detect, whether an individual finds it to be a trustworthy or untrustworthy partner. This thesis describes the design, implementation, and validation of a computational model for recognizing interpersonal trust in social interactions. By first investigating unconscious signals like synchrony, mimicry, and gestural cues, we identify how and which of these nonverbal behaviors are predictive of later cooperative, or trusting, outcomes. We then design a probabilistic graphical model around those resulting predictive cues, and through state-of-the-art 3D motion capture technology, we automate this recognition of trust by tracking those trust-related behaviors.

1.2 Thesis Overview

This thesis is structure with the following:

- **Chapter 1** *Introduction*: We began by stating the overarching goals and motivations of this thesis.
- **Chapter 2** *Background*: We identify and review synchronous movement, mimicry, and gestural cues as nonverbal behaviors with strong relations to trust and rapport.
- **Chapter 3** *Study 1 - Identifying Trust-Related Nonverbal Behaviors*: By leveraging a pre-existing dataset of people socially interacting in the "wild", we investigate how those three nonverbal behaviors are predicative of later cooperative or uncooperative outcomes, and we identify 7 gestural cues of import.
- **Chapter 4** *Modeling Interpersonal Trust from Predictive Cues*: We test and train a Hidden Markov Model (HMM) to recognize whether an individual will have higher or lower levels of trust for their novel partner by observing how these 7 gestural cues unfold in the social interaction.
- **Chapter 5** *Study 2: Capturing 3D motion for Gesture Recognition*: By capturing 3D motion data using the Microsoft kinect, we record how participants move in a social interaction and use this data to train and test 4 support vector machines (SVMs) to recognize the 7 gestural cues of import.
- **Chapter 6** *Validation of System*: By combining both the gesture recognition and trust recognition systems, we demonstrate in a final test how accurately we can autonousmously assess whether an individual is willing to behave cooperatively or uncooperatively with their novel partner.
- **Chapter 7** *Conclusion*: We end with a summary of the contributions of this work and provide future directions for continued work.

Chapter 2

Background

2.1 Overview

Pulling work from the field of human social psychology, embodied conversational agents (ECAs), and human-robot interaction (HRI), we explored various nonverbal behaviors that have been found to be strongly related to interpersonal trust and rapport. And synchronous movement, mimicry, and gestural cues were identified as important unconscious behaviors to further investigate when designing a computational model of trust.

2.2 Synchronous Movement

Synchronous movement, defined by Bernieri [1988], is the precise timing and coordination of movements to coincide with the timing or rhythm of the movements of another - timing that is synchronized or simultaneous. And this intimate coordination has been found to be highly correlated with rapport. In observing student-teacher interaction dyads, raters perceived a higher degree of synchronous movement in dyads that also self-reported higher positive rapport [Bernieri, 1988]. And by having participants rock in chairs together [Valdesolo et al., 2010], walk in time [Wiltermuth and Heath, 2009], or tap fingers in unison [Valdesolo and DeSteno, 2011], researchers can induce a sense of perceptual sensitivity through this "muscular bonding" which leads participants to have increased success in joint-action tasks, have higher levels of cooperation, and even have increased feelings of compassion. And beyond human-to-human social interaction, robots can also elicit different behaviors from people depending on their synchronous versus asynchronous movements. In a dancing interaction, Keepon, a small squash-and-stretch robot, was more successful in engaging children to dance in coordination with the robot when its movement were synchronized to the rhythmic music [Michalowski et al., 2007].

There is compelling evidence that synchronous movement, as either an unconscious

or conscious mechanism, functions as a subtle nonverbal behavior that can lead to enhanced coordination and cooperation. But, research looking at intentional synchrony (like conscious rocking or tapping), although inspiring, does not reveal the resulting nature of unconscious synchrony that occurs naturally in social interactions. As such, we investigate the relationship between *unconscious* synchronous movement and trust between dyads in our study while also utilizing more robust measurements for synchrony through computer vision algorithms.

2.3 Mimicry

We mimic each other in all sorts of ways - from emotions and facial expressions to appearance and even grammar [Chartrand et al., 2005]. But in particular, mimicry in terms of postures and gestures have been identified as strongly linked to trust, liking, and rapport. This "chameleon effect" of behavioral mimicry has been found to facilitate trust in negotiations which resulted in better pay-offs for the intentionally mimicking negotiator [Maddux et al., 2007]. Also, confederates (human actors) who intentionally mimicked the postures and movements of participants are rated to be more likable than those who did not mimic [Chartrand and Bargh, 1999]. And when interacting with an agent in a virtual environment, participants similarly rated mimicking virtual agents as more likable, persuasive, and engaging than non-mimicking ones [Bailenson and Yee, 2005; Gratch et al., 2006].

There is an abundant source of research investigating mimicry. However, there are some unanswered questions when trying to understand how mimicry is predictive of trust. One big question is: which mimicry behavior is worth observing when predicting trust? Will any type of mimicry work? Unfortunately, previous research that used intentionally mimicking agents, either through a confederate or a virtual robot, has a few shortcomings when trying to answer this question.

Both of the virtual agent studies in [Bailenson and Yee, 2005; Gratch et al., 2006] had a two-condition setup, where in the responsive condition, the agent would mimic the participant's posture shifts, head movements, gaze behaviors, shakes, and nods. In the unresponsive condition, the agent would playback a prerecorded random script or that of a prior interaction. The unresponsive condition however would lead to a robot behavior that was completely independent and blind to the participant's behavior, making it difficult to attribute the findings to be a direct result of mimicry or of contingency. And when using human confederates as intentional mimickers, researchers would instruct the confederate to mimic the posture and movement of the other participant, and unfortunately they do not detail what behaviors the confederates decided to mimic [Maddux et al., 2007].

Beyond intentional or conscious-type mimicry, we take a step closer into the context we are more interested in, *unconscious* mimicry, through Chartrand and Bargh [1999]'s and Lakin and Chartrand [2003]'s work. Again, confederates are used, but in a more clever way. By having the confederates either shake their foot or rub their face, researchers looked to see if the unknowing participant would then unconscious mimic those particular behaviors . These passive simple behaviors (that are out of any context or meaning) are good indicators for assessing differing amounts of mimicry, but those particulars cues are not necessarily indicative of higher or lower levels of trust. By observing unconscious mimicry occurring between two participants (and no confederates), we hope to gain a better insight into meaningful mimicking behaviors and how they can be predictive of trust.

2.4 Gestural Cues

Recent research has found a set of gestural cues, leaning-back, crossing-arms, face-touching, and hand-touching, that when taken together, are predictive of uncooperative behavior or lower levels of trust [DeSteno et al., 2011]. And in the same way, Trout [1980] found that when dyads both assumed a forward-leaning posture, they are perceived as having higher degrees of rapport. In a much broader sense, Tickle-Degnen and Rosenthal surveyed many studies to find gestural cues that were related to general positive feelings (like warmth, empathy, understanding, genuineness, friendliness, liking), and reported moderate-to-large relationships with forward leans, smiling, nodding, direct body orientation, and uncrossed arms [Tickle-Degnen, 1990]. We take inspiration from these studies and hope to reconfirm and extend these already associated trust-related gestural cues.

2.5 Trust Recognition Systems

To the best of our knowledge, applications of trust recognition systems currently only exist in the context of assessing trust and reputation of buyers and sellers in online communities. Many different computational models exist in deciding whether an online service and product is trustworthy or reputable in the electronic marketplace (see [Pinyol and Sabater-Mir, 2011] for an extensive survey). By observing features like transaction histories, consumer ratings of sellers, and peer-to-peer recommendations, online services like Amazon and EBay utilize these computational models to provide recommendations to their users [Sen and Sajja, 2002].

Sandy Pentland's sociometric badge, in similar spirits, observes the honest signals exhibited from our voice, body movements, and proximity to others to understand human behaviors and networks [Pentland, 2008]. By means of a microphone, two

accelerometers, and an IR transceiver, this sensor-based modeling of human communication can represent influence between individuals and is capable of predicting turn-taking behaviors in group discussions with 77% accuracy [Pan et al., 2011]. And just through prosody and vocal tone alone, Soman [2009] can predict, with 87% accuracy, the outcomes of job interviews by observing the activity, engagement, mirroring, and emphasis of speech between an interviewer and interviewee.

With no prior art in interpersonal trust recognition systems, we take some initial strides in computationally modeling the dynamics of nonverbal behaviors on interpersonal trust in social interactions. And in this thesis, we present a novel trust recognition system that is capable of predicting, with 94% accuracy, whether an individual finds their partner to be trustworthy or untrustworthy.

Chapter 3

Study 1: Identifying Trust-Related Nonverbal Behaviors

3.1 Overview

By leveraging pre-existing social interaction datasets, we investigate how synchronous movement, mimicry, and gestural cues, three nonverbal behaviors identified in literature as related to trust and rapport, can accurately predict whether a participant is willing to behave cooperatively or uncooperatively with their novel partner. The data from this study was previously collected by David DeSteno's Social Emotions Group at Northeastern University. This experiment produced valuable interaction data, which consists of the raw videos of the experiment, video-coded annotations of the participant's behaviors, and trust and liking measurements.

Figure 3-1: Lab room setup for Study 1

The study consisted of two parts. Participants began by engaging in a social interaction with another random participant for 5 minutes. This part of the study was held in a lab room, where participants were seated at a table as shown in Figure 3-1. On the table were slips of paper with conversation topics that ranged from "Where are you from" to "What do you like about Boston." They were given these conversation topics as suggestions but were encouraged to discuss anything minus the experiment

itself. Around the room, three time-synced cameras captured the frontal-view of each participant along with a side-view of the participants (the perspective that is shown in Figure 3-1). For the second half of the experiment, each participant individually played the "Give Some Game" (explained in Section 3.2.1) in separate rooms and also answered some questionnaires.

A total of 41 dyadic interactions, or 82 individuals, participated in Study 1, and 5 dyads were excluded from the analysis for reasons found in Table 3.1, leaving a total of 36 dyads or 72 participants. The pool of participants were undergraduates attending Northeastern University in Boston, Massachusetts. 72% of the participants were female and 28% male. The dyads were randomly assigned yielding 20 female-female pairs, 4 male-male pairs, and 12 mixed pairs.

Table 3.1: Dyads excluded from Study 1 Analysis

# Excluded	Reason	Comments
2	Already knew each other	affected the game's outcome
1	Participant confused about GSG game	affected the game's outcome
1	In a hurry and rushed experiment	quality of interaction hindered
1	Did not take experiment seriously	quality of interaction hindered

3.2 Evaluation

When trying to understand the relationship between trust and synchronous movement, mimicry, and gestural cues, we first need to define how to measure and quantify these items appropriately for our social interaction context. We behaviorally measured how much an individual trusted their partner through an economic cooperative game. A broad set of gestures were video-coded in order to know how much and when participants displayed particular gestural cues. And after carefully choosing an appropriate time-lag interval, we ran a time-lag analysis, which calculates how often a certain event is followed by a target event within a specified time-interval, for all the coded behaviors to measure how much, when, and which gestures participants mimicked. And by utilizing a background subtraction algorithm, we were able to detect how many times a dyad would have a synchronized event. And these events were counted up, and the sum represented their synchronous movement score.

3.2.1 Measuring Trust

A participant's judgement of trust toward their novel partner was behaviorally measured through a Give Some Game (GSG) [Van Lange and Kuhlman, 1994]. The Give Some Game is similar to the Prisoner's Dilemma game in that it represents a choice between self-interest behavior and cooperative-behavior [DeSteno et al., 2011]. The outcome of the game depends on what each individual player decides to do on their own. The game starts with each player possessing 4 tokens. Each token is worth \$1 to the player, but it is worth \$2 if the player decides to give them away to their partner. And just like the player, their partner has to make this decision as well, and they are not able to communicate a strategy beforehand. An example outcome:

> *The player decides to keep 3 tokens and give 1 token away. That means the player will definitely be awarded \$1*3 = \$3. Now if the* ***partner*** *decides to keep 1 token and give away 3 tokens, then the player will get an additional*

*$2*3 = $6, for a total of $9. And the partner will be awarded $1*1 = $1 for the token he kept, plus $2*1 = $2 for the token the other player gave, for a total of $3.*

For maximum individual payoff, one player needs to **keep** all the four tokens, while the partner **gives away** all his tokens, leaving him with $0 and the other player walking away with $12. For maximum communal benefit, both of the players need to give away all the tokens to each other, awarding each player $8.

The number of tokens a participant decided to give their partner represented how much he/she trusted the partner to play cooperatively. With a rating scale from 0 to 4, this behavioral measure more accurately evaluates feelings of trust over traditional methods like questionnaires.

3.2.2 Measuring Gestural Cues

For this exploratory study, the Social Emotions Group chose a broad spectrum of gestural cues that were found in literature as well as common conversational gestures observed in their previous studies. The following gestures were video coded using the Noldus ObserverXT software: smiling, laughing, leaning-forward, leaning-backward, making eye contact, looking away, arms-on-table, arms-in-lap, crossed-arms, open-arms, moving hands as a conversational gesture, hair-touching, face-touching, hand-touching, and body-touching (also listed in Table 3.2). The two video-coders that manually annotated these gestures had high to moderate inter-coder reliability with Cohen's $\kappa = 0.572$ (moderate agreement) and Spearman's $\rho = 0.925$ (high agreement).

3.2.3 Measuring Mimicry

By definition, unconscious behavioral mimicry occurs when one person unknowingly imitates or repeats the behavior of another person. In terms of timing, a valid mim-

icked action is one that begins *after* a persons emits the behavior but *before* losing a sense of contingency. Unfortunately, previous work rarely define their lag interval for mimicry. In their study, Maddux et al. [2007] told participants to *"mimic the mannerisms of your negotiation partner to get a better deal... However, they say it is very important that you mimic subtly enough that the other person does not notice what you are doing."* Here they non-deterministically define when to mimick those gestures. In contrast, Bailenson et al. specifically define the time-lag with 1, 2, 4, and 8 seconds and found a 4-second-delay to be optimal in minimizing detection of mimicry and maximizing interaction responsiveness with their virtual avatar [Bailenson et al., 2004; Bailenson and Yee, 2005]. However, in both these works, they used agents that intentionally mimicked, which is a behavior not examplary of what is seen in the "wild." Thus, when investigating mimicry that occurs **unconsciously**, we looked at the raw interaction videos to gain a better idea of the appropriate time-lag interval for natural mimicry.

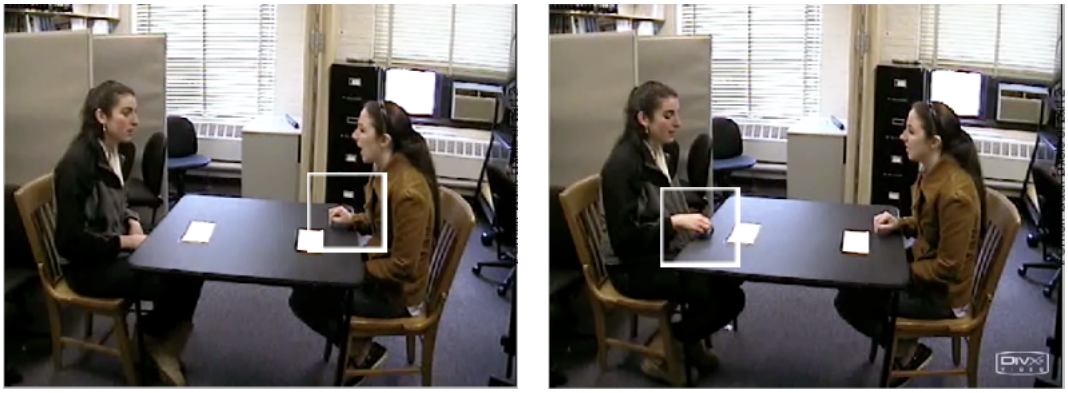
(a) frame = 0 | PartA (left) scratches nose (b) frame = 36 | PartB (right) scratches eye

Figure 3-2: PartB mimicking PartA's face-touch after 1.2 seconds.

[Case 1: Face touch mimicked after 1.2 seconds] *"You rub your chin-darned if my hand doesn't gravitate toward my chin as well"* [Chartrand et al., 2005]. A very common mimicked gesture is to touch the face, whether to scratch the nose or rub the

chin. In Figure 3-2, participant A (partA) scratches her nose, and 36 frames later, participant B (partB) scratches her eye. Recording at 30 fps, the time-lag to mimic this gesture is 1.2 seconds.

[Case 2: Gesture on table mimicked after 4.5 seconds] Some gestures are slower to mimic than others. For example, leans (forward and back) are usually mimicked after a 1-second-delay. In Figure 3-3, PartB is explaining her undergraduate course track while gesturing with her hands on the table, and 4.5 seconds later PartA explains her progress using the same gesture.



(a) frame = 0 | PartB gestures on table (b) frame = 134 | PartA gestures on table

Figure 3-3: PartA mimicking PartB's table gesture after 4.5 seconds.

[Case 3: Smile mimicked after 333 milliseconds] Other gestures are very quick like head nods and smiles. In Figure 3-4, PartA quickly mimics PartB's smile only after 333 milliseconds.

Defining the time-lag interval for mimicry

By observing different types of mimicry occurring naturally in a social interaction, we can approximate through this empirical evidence that non-conscious mimicry has a time-lag interval of 333 milliseconds to 4.5 seconds. Actions that were mimicked later

(a) frame = 0 | PartB (bottom) smiles (b) frame = 10 | PartA (top) smiles

Figure 3-4: PartA mimicking PartB's smile after 333 milliseconds.

than 4.5 seconds (like 8 seconds) could have a loss of connectedness or contingency, like how participants in [Bailenson et al., 2004] had a difficult time detecting mimickers (that mimicked after 8 seconds) even though they were consciously trying to find them.

Mimicry Score

For every dyad, we ran a time-lag analysis of [.333, 4.5] seconds for all the coded behaviors. Each dyad then received a mimicry score which represented a sum of how many times participant A mimicked participant B's gestures and vice versa. However, we decided to remove eye behaviors (eye contact, looking away) as possible mimicked gestures because their quick movements made coding very difficult and unreliable. Also as seen in Table 3.2, the sheer number of their mimicked occurrences (800+) dwarfed other mimicked behaviors like laughing (90+), causing mimicked eye behaviors to have too large of an influence on the mimicry measure. In the appendix,

table A-1 shows the final mimicry score for each dyad.

Table 3.2: Summary of mimicry instances seen across the 36 dyads of Study 1

Gesture	# of Mimicry Instances
Smiling	313
Laughing	98
Lean Forward	11
Lean Backwards	4
Eye Contact	810
Looking Away	999
Arms on table	37
Arms in Lap	37
Crossed Arms	6
Open Arms	3
Conversational Gesture	67
Hair touch	3
Face touch	30
Hand touch	52
Body touch	67
Head Nod	356
Head Shake	99

3.2.4 Measuring Synchronous Movement

Synchronous movement is temporally different from mimicry. Synchronous movement by definition is movement that occurs simultaneously with zero to negligible delay. Traditionally in the field of social psychology, researchers would prime participants by having them perform synchronous acts together like walking in time [Wiltermuth and Heath, 2009], rocking in chairs [Valdesolo et al., 2010], or tapping fingers to music [Valdesolo and DeSteno, 2011] in order to study the effects of synchronous activity. Or by using Likert scales, researchers would measure synchronous movement by having judges rate the level of simultaneous movement, tempo similarity, and smoothness of interacting participants [Bernieri, 1988; LaFrance, 1979].

In an effort to have a more reliable and quantified synchronous movement measure,

we used a simple background-subtraction computer vision algorithm to capture the overall movement of the participants and applied some smart thresholding to obtain a synchronous movement score for every dyad.

Background Subtraction

Although the simplest algorithm for background subtraction, the frame difference method is arguably one of the most robust techniques in detecting a moving foreground. One of its major advantages is its quick adaptability in removing the background since every calculation only depends on the current and previous frame; this also helps in removing background noise such as waving tree or in our case illuminated window blinds [Benton, 2008].

$$I_{ij}(t) = \quad | \, frame_{t-1} - frame_t \, | \tag{3.1}$$

The algorithm consists of taking the absolute difference between the grayscale pixels in the current frame from the previous frame (Equation 3.1). Figure 3-5 shows how the algorithm reveals the moving foreground as the nonblack pixels represent the movement of a participant adjusting in his chair.

Figure 3-5: Background subtracted frame of a participant moving in his chair

Thresholding

At every frame difference in the 5 minute interaction, we counted the number of nonblack pixels on the left half of the image to represent the overall movement of participant A ($Activity_A$) and the right half to represent participant B ($Activity_B$). We placed a lower-bound threshold, or cutoff, on this *Activity* so that we would only capture significant movement and discount any noise. We empirically found this appropriate threshold by looking at the background subtracted videos.

[Case1: Threshold at 425] In Figure 3-6, participant A is adjusting in his seat as participant B is gesturing with her hands. The movement of seat adjustment peaks with an activity level of 425 nonblack pixels, while hand gesturing peaks at 950, making 425 a good value to capture both of these motions.

Figure 3-6: Activity levels of participant A (left) adjusting in his seat and of participant B (right) gesturing with her hands

[Case2: Threshold at 250] In Figure 3-7, participant A is leaning-forward as participant B is reaching to touch her hair. The movement of leaning-forward is capture with activity levels above 250 nonblack pixels, while the hair-touch action peaks at 1700, making 250 a good lower-bound threshold value.

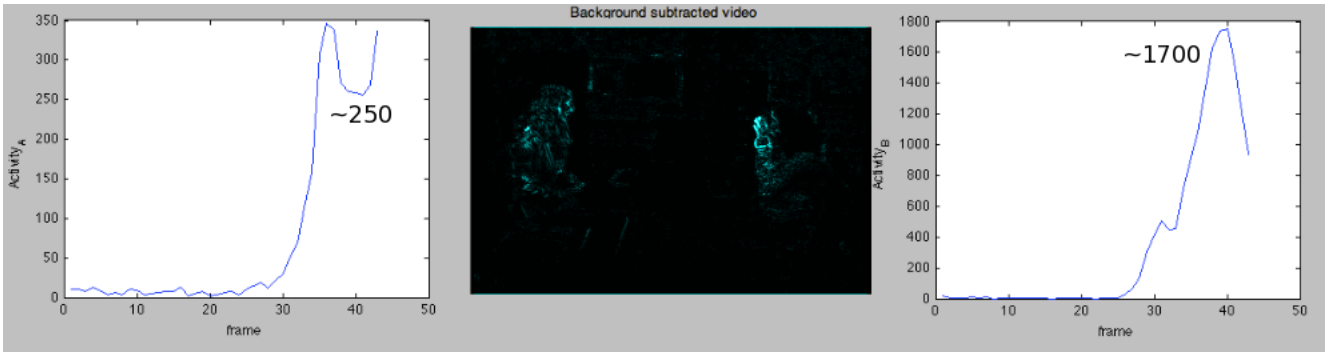


Figure 3-7: Activity levels of participant A (left) leaning-forward and of participant B (right) reaching to touch her hair.

Synchronous Movement Score

After considering Case 1 and Case 2, a lower-bound threshold of 250 nonblack pixels ($P(t)$) was the minimum required to be considered a significant change in motion between two consecutive frames. And over the entire 9,000 frames of interaction, the function $Activity(t)$ represents how much a participant is moving at a particular frame (Equation 3.2).

$$I'_{ij}(t) = \begin{Bmatrix} 1, & I_{ij}(t) > 0 \quad \forall_{ij} \\ 0, & otherwise \end{Bmatrix} \tag{3.2a}$$

$$P(t) = \sum_{ij} I'_{ij}(t) \tag{3.2b}$$

$$Activity(t) = \begin{Bmatrix} P(t), & P(t) \geq 250 \\ 0, & otherwise \end{Bmatrix} \tag{3.2c}$$

By multiplying the activity of participant A ($Activity_A(t)$) to the activity of participant B ($Activity_B(t)$), we can witness when a dyad ($Activity_D(t)$) exhibits simultaneous activity. For example, Figure 3-8 shows such synchronized events of dyad 66 over the entire social interaction. And as evidenced by Figure 3-9, we can verify that these events (like Event A, B, and C) are indeed representative of synchrony occurring between the two participants.

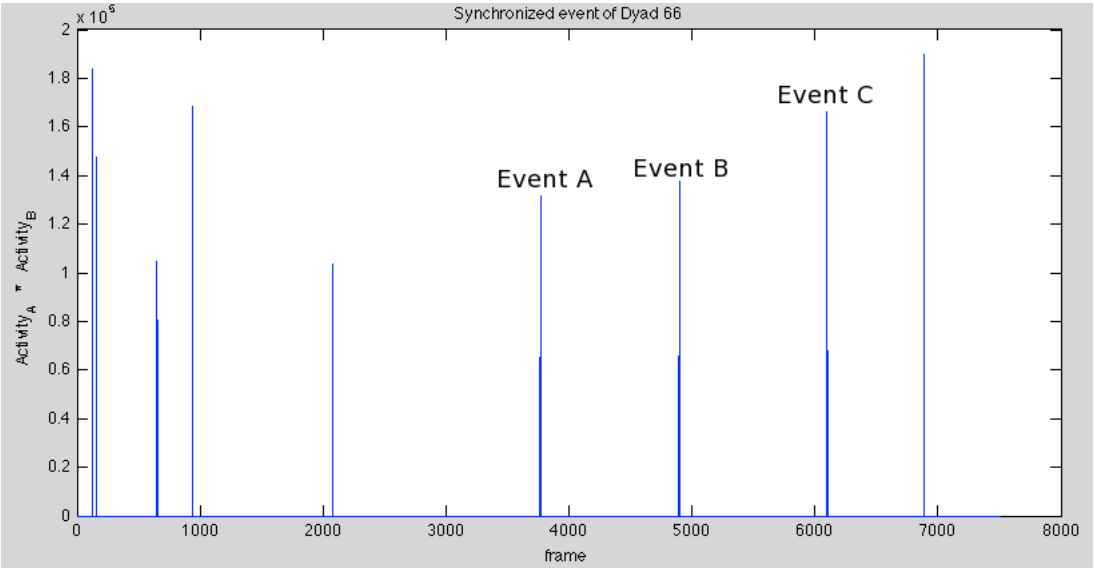
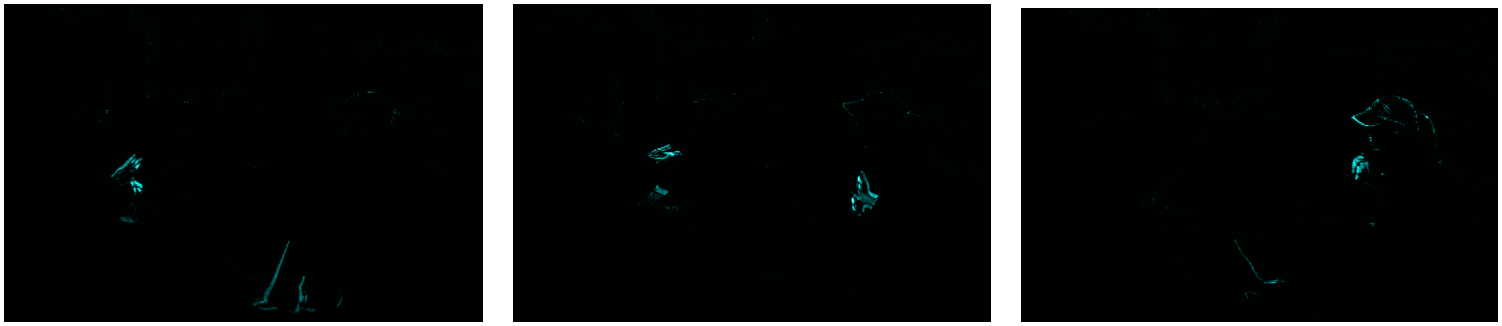


Figure 3-8: Synchronized events of dyad 66 over the entire 5 minute social interaction with labeled events A, B, and C.



(a) Event A (b) Event B (c) Event C

Figure 3-9: Verification that synchronized events in Figure 3-8 are indeed representative of synchronous movement occurring in the interaction. In Event A, PartA hand gestures as PartB kicks foot. In Event B, both are hand gesturing. In Event C, PartA kicks foot as PartB nods.

And finally, by counting the total number of frames where such synchronized events occurred, we obtain a reliable and quantified synchronous movement score for every dyad (Equation 3.3). Table A-1, in the appendix, shows the final scores.

$$Activity_D(t) = Activity_A(t) * Activity_B(t) \tag{3.3a}$$

$$Activity'_D(t) = \begin{Bmatrix} 1, & Activity_D(t) > 0 \\ 0, & otherwise \end{Bmatrix} \tag{3.3b}$$

$$Sync = \sum_t Activity'_D(t) \tag{3.3c}$$

3.2.5 Method of Analysis

Linear regressions fitted using the least squares approach were used as the primary method of analysis, and we report the slope value b along with its significance level p. Since a regression analysis attempts to describe the response (Y) as a function of a dependent variable (X), it inherently assumes there is a one-direction causal effect from the dependent variable to the response variable. As such, we can find whether synchronous movement, mimicry, or gestural cues can be predictive of trust. And in cases where the direction of the relationship of two independent variables is unknown, we used a zero-order correlation analysis to determine the statistical significance of the correlation coefficient r between the two variables. The correlation coefficient is reported with the degrees of freedom in parentheses along with its significance level p.

3.3 Results

3.3.1 Synchronous Movement

An additional three dyads had to be removed for only the synchronous movement analysis (n = 33 dyads). Two were due to extraneous disturbances causing poor performance from the background subtraction algorithm. In one case, a participant was wearing a high contrast shirt that made any slight movement become rated as large ones. And in another, the sunlight at that time period scattered onto the window binds in such a way that its shifting illuminance interfered with the dyad's sync score. The third dyad was simply due to delayed video recordings.

And surprisingly, *synchronous movement, in a linear regression analysis, was found to not be predictive of trust* [b = 0.0002, p = 0.955]. However, in a zero-order correlation analysis, *synchronous movement and mimicry were significantly positively correlated* [r(31) = 0.474, p $<$ 0.005]. When dyads mimicked each other more, they also moved more synchronously. We were concerned that perhaps the synchronous movement score was just a residue of mimicked actions, i.e. when a participant nods and the other immediately follows, there could be a duration where they are both nodding away. And such an event would increase both the mimicry score as well as the sync score. Pursuing this further, we re-calculated the mimicry score for all the dyads but with a time-lag-analysis of [2, 5] and [4, 8] instead of [.333, 4.5]; by looking at slower mimicked actions, we can avoid this double counting. And even still, we found synchronous movement and mimicry to be significantly positively correlated with [r(31) = 0.396, p $<$ 0.023] [r(31) = 0.438, p $<$ 0.011], respectively.

Although our primary focus is understanding trust, we also asked the participants how much they liked their partner with a 7-point Likert scale question. And going against previous findings [Valdesolo and DeSteno, 2011], *synchronous movement was*

not found to be predictive or correlated with dyad liking (calculated with average of individual reports) in a linear regression and a zero-order correlation analysis. We also looked to see if there were any differences in synchronous movement in same-gender versus mixed-gender dyads. Using an unpaired 2 tailed t-test of unequal variance, *we found that same-gender dyads synchronized significantly more than mixed-gender dyads* [p = 0.0165, $\mu_{mm,ff} = 144$, $\mu_{mf} = 64$] (see Figure 3-10).

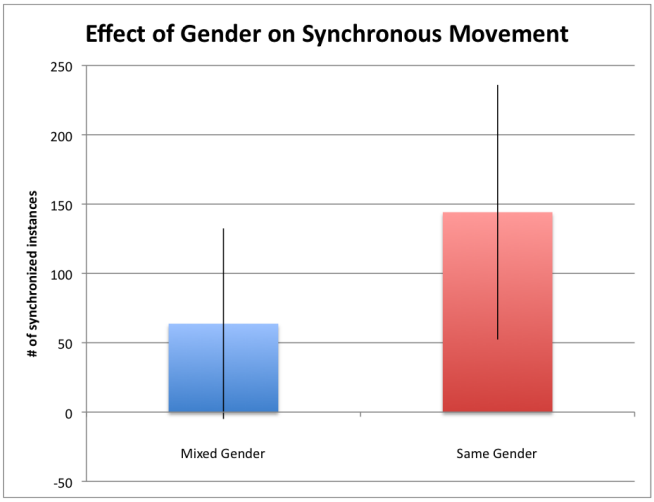


Figure 3-10: Same-gender dyads synchronized statistically more than mixed-gender dyads [p = 0.0165, $\mu_{mm,ff} = 144$, $\mu_{mf} = 64$]

3.3.2 Mimicry

In a linear regression analysis, *mimicry was found to not be predictive of trust* [b = -0.003, p = 0.540]. Going against previous results in literature, this finding is thoroughly discussed in Section 3.4 below. We also tried grouping different types of mimicked actions, hoping that perhaps a constellation of mimicked actions like leaning-back, crossings-arms, and hand-touching could be associated with lower levels of trust. Unfortunately, no such sets were found. However, agreeing with previous literature, *we found that mimicry was predictive of liking.* When a participant mim-

icked their partner more, they rated their partner as more likable [b = 0.038, p < 0.014], and so did the partner in return [b = 0.034, p < 0.027]. And also at the dyad level, the more that dyads mimicked, the more they liked each other (average of ratings) [b = 0.021, p < 0.025].

3.3.3 Gestural Cues

Reconfirming the results found in [DeSteno et al., 2011], the set of four cues, face-touching, crossing-arms, leaning-back, and having arms on the table (which in coding usually co-occurs with hand-touching), when taken together were predictive of distrust. Through a linear regression analysis, we found that the more a participant exhibited these disengaging cues, the fewer tokens they gave their partner [b = -0.020, p < 0.040].

Although these negative cues were strongly predictive of lower levels of trust, we wanted to identify some positive cues that could be associated with higher levels of trust. Since these negative cues were very strongly associated with negative trust, it was difficult to find cues that were positively associated with trust when using the entire dataset (n = 72 participants). In an effort to tease out some of these positive cues, we randomly took half of participants (n = 36) and found a set of three cues, leaning-forward, arms-in-lap, and having an open-arms pose, that when taken together, are to be predictive of higher levels of trust [b = 0.047, p < 0.031].

Table 3.3: Set of positive and negative gestural cues that are predictive of trust.

Set of positive cues	Set of negative cues
Lean forward	Lean backward
Open arms	Crossed arms
Arms in lap	Hand touches
	Face touches

3.4 Discussion

3.4.1 Synchronous Movement

Although synchronous movement was not predictive of trust, we found that it was significantly positively correlated with mimicry. That is, people who mimicked each other more frequently also move more synchronously in time together. And there is evidence in the literature that supports this finding. Valdesolo et al. [2010] showed how rocking in synchrony enhances an individual's perceptual sensitivity to the motion of another person. In this study, participants were told to either a) synchronously rock together in rocking chairs or b) asynchronously rock. Results showed that participants that were primed with this synchronous rocking increased in their success in a subsequent joint-action activity. And similarly, walking together in step caused groups to have higher levels of cooperation [Wiltermuth and Heath, 2009]. Although the synchrony in these studies was a very conscious process, it suggests that synchrony can be used as a mechanism to enhance a person's feelings of connectivity with another, which could lead into more dynamic interactions like mimicry. In other words, synchrony can cause people to "link up" and form a marker of similar identity that causes the give-and-take dance of mimicry to occur more frequently. This is similar to how increased levels of rapport and interpersonal closeness can cause a person to more frequently mimic another individual [Lakin et al., 2003]. Unfortunately, we cannot confirm a causal relationship between synchronous movement and mimicry from our correlation results, but we can establish that they are closely related behaviors.

Bernieri [1988], by evaluating teacher-student interactions, also found positive correlations between synchronous movement and behavior matching (or pose congruence) using third person ratings. Using more robust techniques for measurement and a stricter definition for behavior matching, we also confirm that these behaviors are indeed very closely related.

We also found that same-gender dyads synchronize more than mixed-gender dyads. This finding suggests that enhanced sense of connectivity, established through similarity in gender, can cause more synchronization between dyads. Prior work has shown synchrony to increase feelings of connectivity, but can increased feelings of connectivity or identity also increase unconscious synchrony? Due to our results, we have reason to believe that this is indeed the case, which establishes a bidirectional relationship between synchrony and connectivity.

3.4.2 Mimicry

Originally, we were surprised that we did not find mimicry to predict trust. However, when we considered the social context of the study in comparison to previous studies in literature, we were encouraged by how our results supported the very versatile nature of unconscious mimicry.

In previous studies, participants were placed in situations where they were actively negotiating to make a deal, turning up their mimicry mechanisms into high gear in order to quickly establish some sort of affiliation. We had hoped that later expectations of cooperation (in playing the Give Some Game) would be enough to cause participants to actively try to mimic one another. But unfortunately, such a disconnect was not enough to induce any significant and meaningful trust-related mimicry behaviors. And in the same light, since synchronous movement is closely related to mimicry, then it is also not surprising that synchronous movement was not predictive of trust for this study.

Although mimicry can be used for building affiliation, it can also be used to build liking. Mimicry has been posed as an adaptation mechanism that helps us fit in amongst others [Lakin et al., 2003]. By mimicking you and becoming more like you,

we seem like we have more things in common. This can lead to greater liking towards each other, which is exactly what we saw in our results. That is, the social situation of our study lead participants to use mimicry as a means to convey liking instead of trust or rapport.

3.4.3 Gestural Cues

We identified 3 gestural cues that were related with higher levels of trust, and reconfirmed 4 gestural cues that were related with lower levels of trust. The following gestures are important for observation when discerning an individual's assessment of their partner's trustworthiness: lean-forward(+), lean-backwards(–), hand-touch(–), face-touch(–), open-arms(+), crossed-arms(–), and arms-in-lap(+). We recognize that the study's friendly and prosocial context sets a default expectation that people will most likely cooperate, making the negative cues of greater predictive value. Although the positive cues were not significantly predictive of trust in its full set analysis ($p < 0.031$ only for $n = N/2$), we are encouraged by the fact that these cues are the direct inverse of the negative cue set, and therefore they should be considered as potentially important positive cues that are predictive of higher levels of trust.

Chapter 4

Modeling Interpersonal Trust from Predictive Cues

4.1 Overview

We previously identified 7 important cues that were predictive of trust: lean-back, lean-forward, crossed-arms, open-arms, face-touch, hand-touch, and arms-in-lap. By observing these cues throughout the duration of the 5-minute social interaction, can we accurately predict how much a participant will trust their partner? Can we build a model that can recognize through the pattern of emitted gestures whether a person trusts another person or not? Believing that not only does the frequency of these predictive gestures matter, but also the sequence in which they present themselves, we decided to use Hidden Markov Models (HMM) to detect trust.

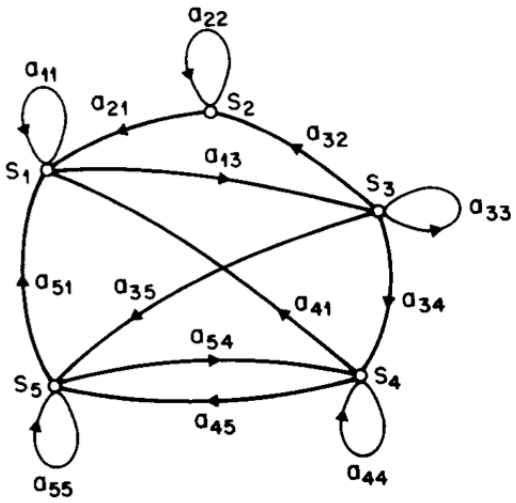


Figure 4-1: Example of a basic Hidden Markov Model where states are represented as nodes and transitions as connecting links between nodes. Image courtesy of [Rabiner, 1989].

HMMs have found much success in modeling processes that unfold over time. In applications that have a temporal progression, HMMs represent these sequences with states $S = \{S_1, S_2, \cdots S_N\}$, where a state at time τ is directly influenced by a state at τ-1 as shown in Figure 4-1 [Duda et al., 2000]. The process being modeled begins at

state $\pi = \{\pi_i\}$, and with every passing observation $O = \{O_1, O_2, \cdots O_T\}$, the HMM either stays in the same state or moves to another state j based on the transition probabilities $A = \{\alpha_{ij}\}$ and the observation probabilities $B = \{b_j(O_t)\}$. When learning an HMM model $\lambda = (A, B, \pi)$, the model's parameters are iteratively adjusted in order to maximize the probability of the model generating the given observation sequence. For a more detailed description of HMMs please see [Rabiner, 1989].

4.2 Training the Hidden Markov Model

Since a participant's decision on how many tokens to give to the partner depends on the interaction (which varies per dyad), one might expect a random distribution of the total number of participants that gave either 0,1,2,3, or 4 tokens. However, only a handful of participants gave 0 or 1 tokens. Giving 2 tokens was the most popular choice, perhaps to its safe nature of giving half. And giving either 3 or 4 tokens was a close second. Due to this distribution, we decided to train one HMM_{low} using data from participants that gave 2 tokens away, which signified low levels of trust, and another HMM_{high} using data from participants that gave 4 tokens away, which signified high levels of trust.

To produce a good generalized model and to avoid overfitting, we need to train with as many examples as possible. As such, we obtain more interaction data from David DeSteno's lab of past studies that were the same in nature as the one described in Section 3.1. In total, we had a collection of 24 training examples for high trust and 50 for low trust. Each training example contained a sequence of observed gestural cues a participant emitted for the entire 5-minute interaction, where the possible observed cues are the $M = 7$ predictive cues. Thus, $O = \{O_1, O_2, \cdots O_T\}$ for our case would look like $O = \{crossedArms, faceTouch, faceTouch, leanBack, \cdots\}$ and T would vary per participant (min = 2, max = 68).

We ran 600 simulations using Kevin Murphy's Bayes Net Toolbox for Matlab [Murphy, 2007], and at every iteration we varied the number of possible states from [1, M] for each of the HMMs and initialized A, B, π with random probabilities.

4.3 HMM Results

A leave-one-out cross-validation was used as the training method to determine the numbers of states for HMM_{low} and HMM_{high}. We ranged both N_L and N_H from [1, M], where M=7 (the total number of predictive cues). Each parameter combination was tested 74 times, where each time we would leave the training data from one participant out and train on the remaining 73. The omitted participant would then be used for testing to determine the parameters' ability to generalize to new data. By means of this leave-one-out cross-validation, our best result with 7 states ($N_L = 7$) for HMM_{low}, and 3 states ($N_H = 3$) for HMM_{high}, had a recognition accuracy of 94% with 70 hits and 4 misses. Figure 4-2 shows the full topology of HMM_{high} with observation probabilities (B) greater than 10%. Figure 4-3 illustrates a reduced HMM_{low} that only shows transitions (A) and observation probabilities (B) greater than 10%. The diagrams illustrate the probability of a gesture being observed at a particular state, and the collection of states and their probabilistic transitions between them that represent the most likely sequence of emitted gestures for that particular HMM. Please see A.2 in the appendix for the exact model parameters.

These models taken together are capable of differentiating, with 94% accuracy (note that chance is at 68% with uneven example sets), whether a participant will give 2 or 4 tokens to their novel partner by observing the sequence of predictive cues they emit.

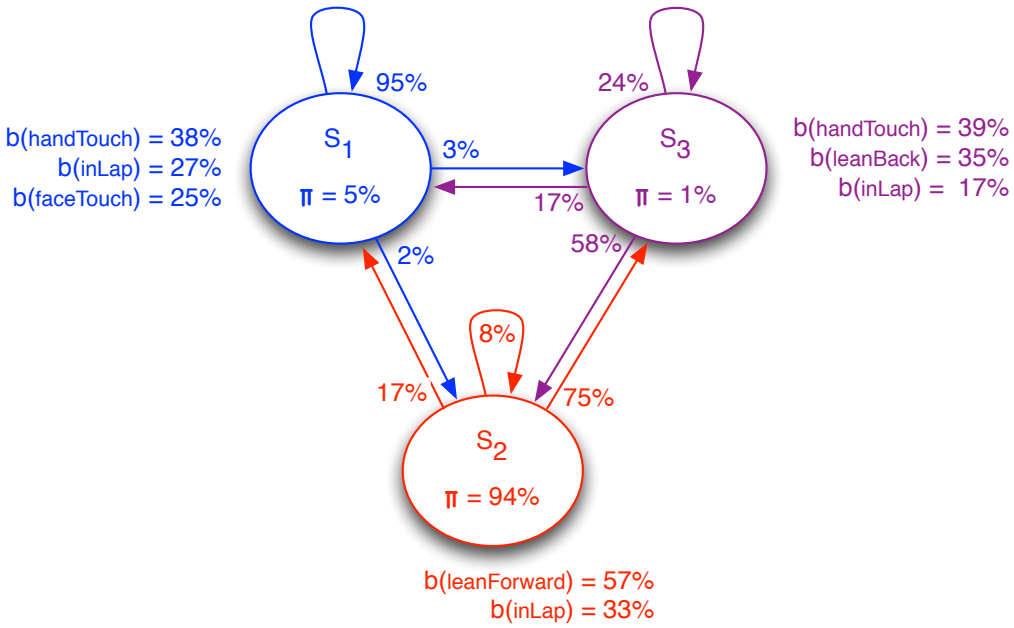


Figure 4-2: Full topology of HMM_{high} with observation probabilities (B) greater than 10%. The most-probable Viterbi path through the HMM state trellis: $S_2S_3S_2S_3S_1S_1S_1$.

4.4 Model Discussion

HMM models and their states usually do not represent the exact truth of a process, instead they are built to maximally differentiate two processes. As such, the HMM_{high} model is not saying "hey this is what trust looks like," but when compared against the other model HMM_{low}, we can extrapolate the major differences in trustworthy versus untrustworthy behavior.

When simulating HMM_{high}, we get an output observation of:

$$leanFwd \rightarrow handTouch \rightarrow inLap \rightarrow leanFwd \rightarrow leanBack \rightarrow inLap \rightarrow leanBack \rightarrow leanFwd \rightarrow leanBack \rightarrow leanFwd$$

When simulating HMM_{low}, we get an output observation of:

$$handTouch \rightarrow leanFwd \rightarrow faceTouch \rightarrow crossedArms \rightarrow inLap \rightarrow$$
$$handTouch \rightarrow faceTouch \rightarrow handTouch \rightarrow inLap \rightarrow handTouch$$

And to make the pattern easier to decipher, we denote $(+)(-)$ as positive and negative cues to form:

$$HMM_{high} = + - + + - + - + -+$$
$$HMM_{low} = - + - - + - - - +-$$

Certainly there is a higher bias to see more positive cues in the trusting model and more negative cues in the lower trust model (for the sake of terminology we will say untrust). But also, there is a more alternating pattern of positive to negative cues in the trust model, while the untrust model shows a back-to-back succession of negative cues being displayed. Thus, as posited previously, the sequences in which we emit these predictive gestural cues also matter.

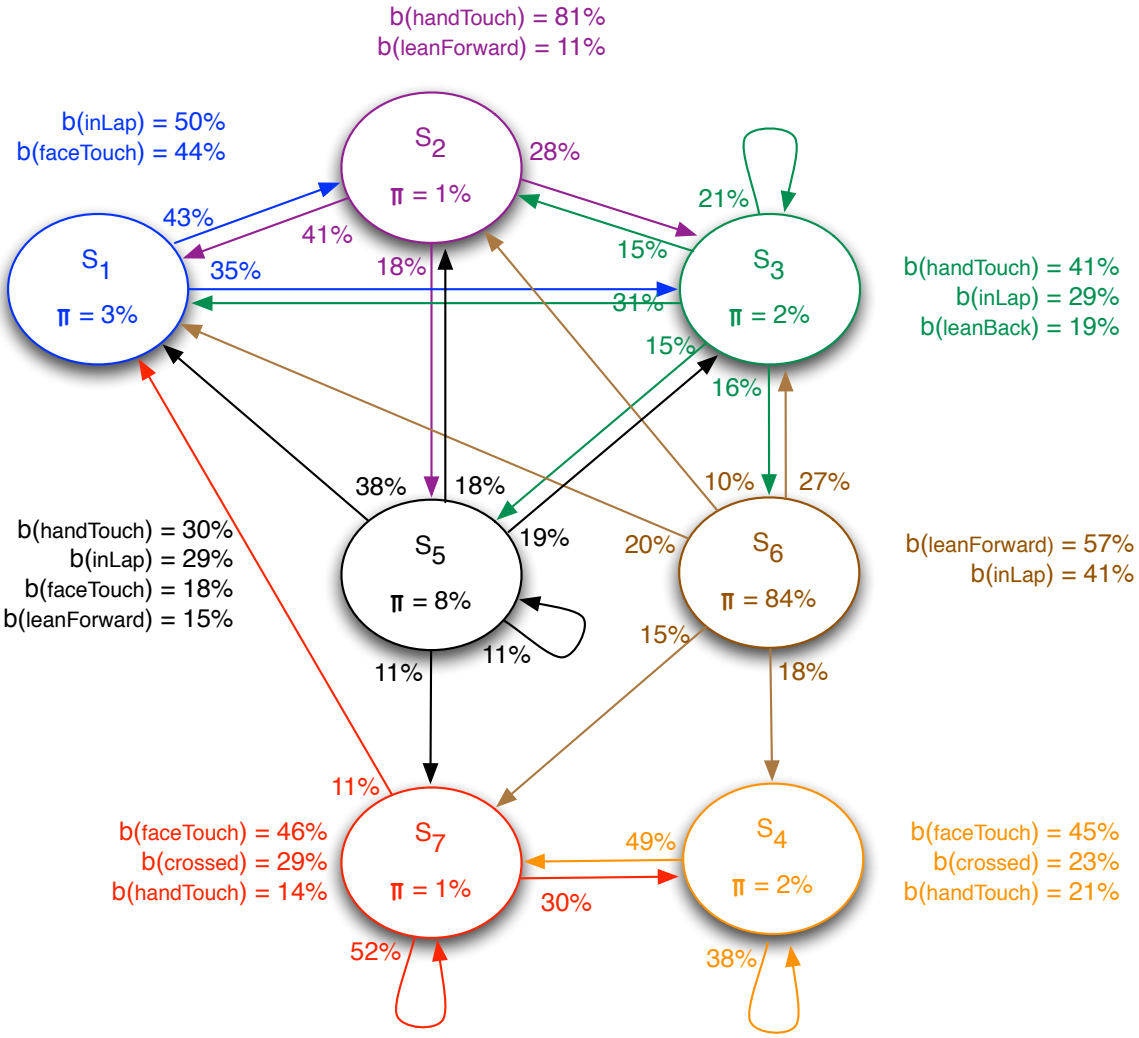


Figure 4-3: A reduced HMM_{low}, showing transitions (A) and observation probabilities (B) greater than 10%. The most-probable Viterbi path through the HMM state trellis: $S_6 S_1 S_2 S_5 S_7 S_7 S_7$.

Chapter 5

Study 2: Capturing 3D Motion for Gesture Recognition

5.1 Overview

We can accurately predict whether a person finds another individual to be trustworthy or untrustworthy by observing a set of informative gestural cues unfold in a social interaction. However, the method in which we obtained the sequence of emitted gestures was through rigorous hand coding. For a robot to determine how much an individual trusts the robot, it needs to recognize these gestures autonomously. To model these gestures, we need a full-body digital perception of how people move in a social interaction. By using 3D motion capture technology, we can track the body movements of people. And through machine learning and gesture recognition algorithms, we can detect when these nonverbal cues are being communicated.

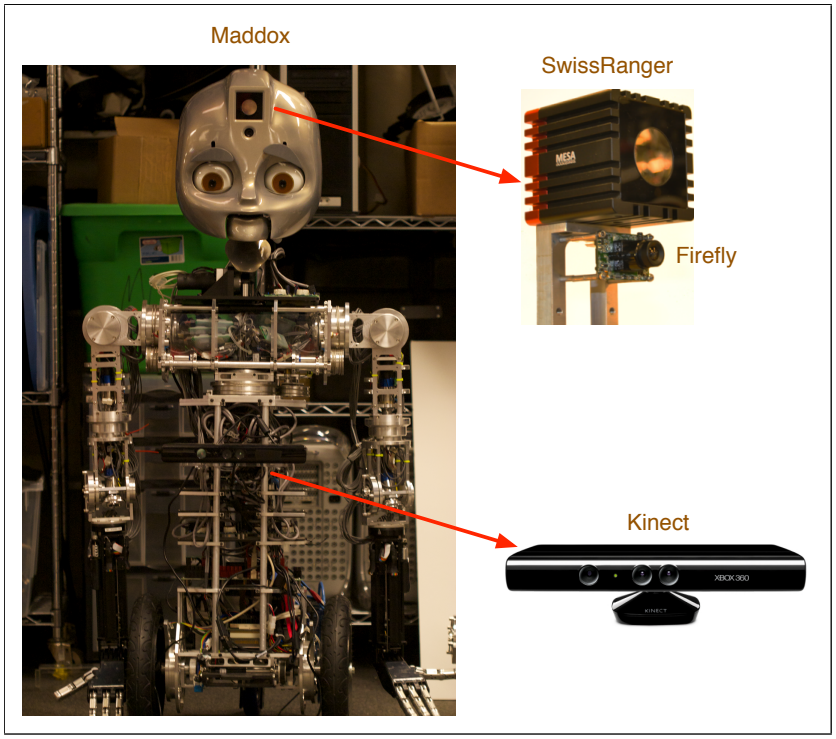


Figure 5-1: The Personal Robot Group's humanoid robot Maddox equipped with swissranger, firefly and kinect sensors.

5.2 Technology for Motion Capture

Our humanoid robot Maddox shown in Figure 5-1 is equipped with two 3D sensors: Microsoft's kinect and MESA's swissranger. The kinect provides a depth map and RGB image map of a scene, and with OpenNI's NITE algorithm we can track a person's movements using their skeleton-tracking algorithm (pipeline illustrated in Figure 5-2).

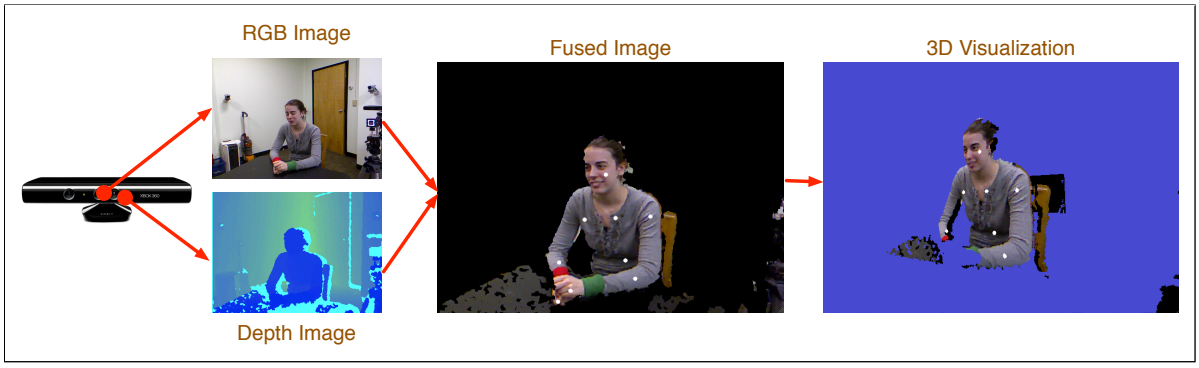


Figure 5-2: The kinect captures depth and RGB images, which are fused with the skeleton data and visualized in 3D.

The swissranger in conjunction with PointGrey's firefly camera also provides a 3D depth map and RGB image, and the Googletracker (which is not commercially available) can track various features of a person's face (pipeline illustrated in Figure 5-3).

The information from both these sensors might seem redundant, but the kinect, which is fixed on Maddox's torso, is appropriate for detecting a whole body with its wide field-of-view, while both the swissranger and firefly, which are attached to Maddox's pivoted head, is appropriate for zooming onto a person's face with its narrower field-of-view. We also decided to use red and green sweatbands for color-tracking a person's wrists as a system fallback whenever skeleton-tracking fails.

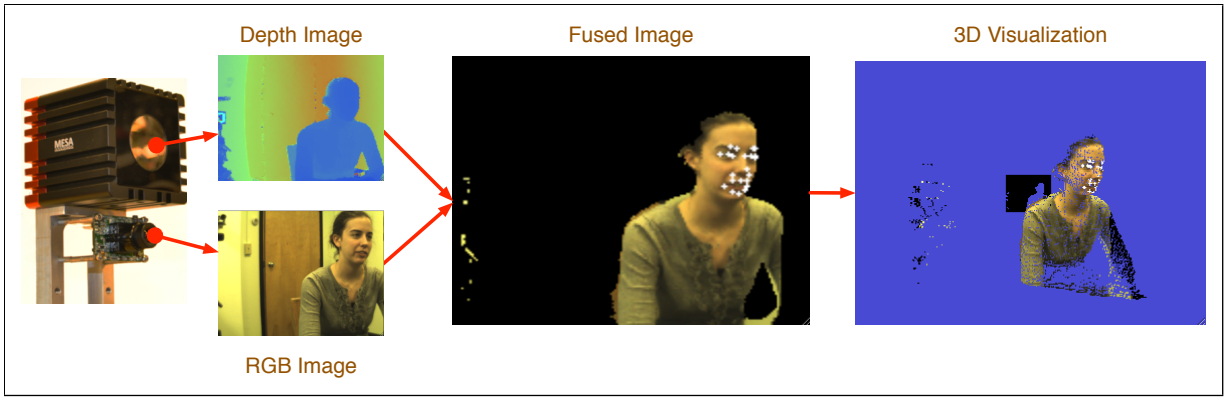


Figure 5-3: The swissranger captures the depth image and the firefly captures the RGB image, which are fused together with the Googletracker face data points and visualized in 3D.

For both the kinect and swissranger/firefly sensors, we used Julian Muñoz [2011]'s vision framework to:

- Calibrate a mapping between the depth and RGB images
- Record the raw depth and RGB values for the study
- Visualize the fusion of the depth, RGB, and skeleton/face-feature data

Using this motion capture system made up of different hardware and software solutions, we were able to capture the face and body movements of the participants for our second study.

5.3 Method

5.3.1 Task

We wanted this study to be the same as the previous study found in Section 3.1 with the only change being the addition of the motion capture system. So again, the task was a 5-minute social interaction followed by the Give Some Game.

5.3.2 Setup

The study was performed in the same lab room as mentioned in Section 3.1. And in addition to the three time-synced cameras necessary for the video coding software, we had two kinects and two swissranger/firefly pair cameras on stands pointed towards the participants as seen in Figure 5-4. On the table, we placed the red and green sweatbands for participants to wear on their wrists. And instead of having the conversation topics available on a piece of paper, we wrote them on a white board adjacent to the participants to avoid having them fiddle with the sheet.

Figure 5-4: Lab room setup for Study 2

5.3.3 Protocol

Once both the participants arrived for the study, they were asked to read and sign both Northeastern's and MIT's consent forms. After giving them some time to read and sign in privacy, we read aloud the following script:

> *The experiment today is part of a joint project between NEU and the Media Lab at MIT. At Northeastern, the lab is interested in studying interpersonal dynamics and how people form impressions of others based on brief social interactions. At MIT, the Personal Robots group works with cutting edge technology developing robots that have the capability of interacting with humans. The lab studies human-robot interactions and they are interested in gathering data to help them improve the technology to enhance such interactions. For the first part of the experiment, we'd like you to please leave all of your things at the computer and have a seat at the table in the center of the lab. In the first part of this experiment, we would like you two to have a conversation for about 5 minutes. But before you start, we are going to ask you to calibrate to some special cameras to capture your interaction. These cameras give us a coarse idea of how you are moving just like the Xbox kinect games.*

We had each participant hold the calibration pose seen in Figure 5-5 for 3 seconds. And continued with the script:

> *Ok, now that we are all set-up we can begin this first part of the experiment. Remember, you two are going to have a conversation for about 5 minutes. We know that all of these cameras may seem distracting, but try to ignore them as best as you can and just act naturally. Feel free to discuss whatever you like; we've provided a brief list of ideas in case you are stuck (see Table 5.1 for the list of questions), but don't feel limited to those conversation*

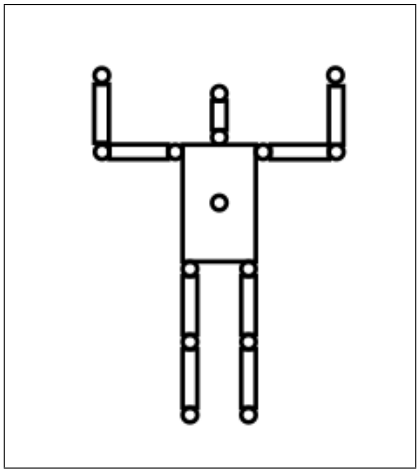


Figure 5-5: Figure illustrates the calibration pose that a user must hold for 3 seconds before skeleton-tracking can begin. Image courtesy of http://www.primesense.com in the NITE Algorithms document.

ideas. However, we do ask that you refrain from discussing the experiment itself. Following your conversation, you will be playing a game together where you have the opportunity to win some money. Any questions at this point? Ok you can go ahead and start the conversation as soon as I leave the room, and I'll be back in about 5 minutes.

Table 5.1: Suggested Conversation Questions

Questions
1) What are your plans for spring/summer break?
2) Have you seen any good movies or read any good books lately?
3) What do you like about living in Boston?
4) What do you like to do for fun?
5) Where are you from?

We then left the room to allow the participants to start their conversations. And after 5 minutes passed, we reentered the room and said the following:

Ok. Time is up. For the next part of the experiment, the two of you will be playing a game called the Give Some Game, and I will need you to complete it in separate labs. So, why don't you come with me and you

can have a seat back at your computer, and another instructor will be here soon.

One participant is escorted down the hall to the second lab. Both the participants were given the follow verbal instructions for the GSG game:

The next part of the experiment is a game called the Give Some Game. In the game, you will have the opportunity to win cash-cash that we will actually hand you at the end of the experiment. You will be playing with the other participant with whom you just spoke. She or he will play the same game at a computer in the lab down the hall. How much money you win will depend both on your decision and on the decision that the other participant makes. In order to ensure anonymity in your decision, you'll be playing the game in separate rooms and you will not see each other at any point during or after the experiment, and you actually leave time-separated, so you shouldn't bump into one another in the hall right after. The computer is going to walk you through the instructions for the game, but feel free to stop and come ask me questions at any point. Remember, you are playing for real money, so it's important that you fully understand the rules of the game. Once you've read through the instructions and examples on the computer, please indicate your decision on the paper we've provided here. I will be back in a few minutes to collect it. Any questions at this point? Ok, go ahead and click the mouse to start.

Once the participant is done with the GSG game and written his/her answer on the answer sheet, we continued with:

While I get your payment in order, there are a few questionnaires that we'd like you to complete on the computer. Go ahead and click your mouse and the computer will walk you through what you need to do. When you're

done, please wait at your computer and I will be back in a minute with your money.

We then interviewed the participants to gauge whether the cameras and wristbands bothered or affected their interactions. And after collecting both answer sheets from the participants, we calculated the earned winning amounts and gave the payments to the participants, which left the lab time-separated.

5.3.4 Participants

A total of 28 dyadic interactions, or 56 individuals, participated in Study 2. The pool of participants were again undergraduates attending Northeastern University in Boston, Massachusetts. 79% of the participants were female and 21% male. The dyads were randomly assigned yielding 18 female-female pairs, 2 male-male pairs, and 8 mixed pairs. 15 participants were excluded from the Trust model analysis for reasons found in Table 5.2. And for the gesture recognition corpus, we had to exclude 3 participants due to kinect software failures that occurred during the experiment.

Table 5.2: Number of participants excluded from Study 2's Trust model analysis

# Excluded	Reason	Comments
5	Played GSG game before	expectations from prior experience
1	Did not understand GSG game	affected the game's outcome
2	Already knew other participant	affected the game's outcome
5	Graduate student from lab	knew too much about the study
1	2nd language english speaker	quality of interaction hindered
1	Significantly fiddled with wristbands	inhibited natural gestures

5.4 Gesture Recognition

5.4.1 Overview

We are interested in detecting the following 7 predictive gestures: lean-forward, lean-back, arms-in-lap, open-arms, crossed-arms, hand-touches, and face-touches. The majority of these gestures are identifiable as a static pose, and some are strictly defined as a pose rather than a dynamic action. For example, crossed-arms can be achieved by various arm movements, but it is primarily identified by the end result of having each arm "crossing" to the other side. Arms-in-lap, open-arms, and hand-touches are also better defined as a static pose. The motions of leaning-forward, leaning-back, and face-touching are unique enough to be defined as a dynamic action, but they can also be identified in their single instance. Unlike a waving gesture where multiple arm swings are necessary to complete a wave, we can identify with just a single frame whether a person is leaned forward or back or touching their face. As such, instead of choosing temporally based models, support vector machines (SVMs) were chosen as our method of gesture recognition.

An SVM model represents a collection of examples mapped in high-dimensional space in such a way as to optimally separate categories with clear boundaries and margins. By mapping the training feature vectors of finite-dimensional space into higher dimensional space, SVMs achieve easier separations using hyperplanes [Duda et al., 2000]. There are four common types of kernel functions used for SVMs: linear, quadratic, polynomial, and radial. We chose to use the radial basis function as suggested in [Hsu et al., 2003] as we have a greater number of instances than features, and we also expect the relation between the class labels and attributes to be nonlinear.

At first intuition, we decided to group gestures that were mutually exclusive into the same SVM model. We began with the following set of SVMs:

Table 5.3: SVM set based on mutually exclusive gestures

SVM_1	SVM_2	SVM_3
Lean forward	Arms in lap	Face touch
Lean back	Crossed arms	Hand touch
	Open-arms	

However, we found that some features were more relevant to particular gestures while irrelevant to others. As such, instead of creating an SVM that tries to differentiate gestures using some good and some bad features that could potentially prevent a SVM from converging to an optimal solution, we further separated the detection of gestures into different SVMs based on the differences in their feature vectors as shown in Table 5.4. These differences were primarily driven by failures in skeleton-tracking, which is explained in Section 5.4.2.

Table 5.4: SVM set based on mutually exclusive gestures and differences in feature sets.

SVM_1	SVM_2	SVM_3	SVM_4
Lean forward	Arms in lap	Open arms	Face touch
Lean back	Crossed arms		Hand touch

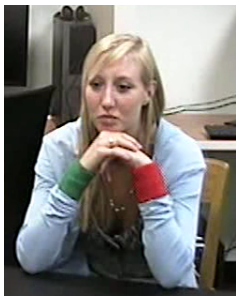


Figure 5-6: A case where hand- and face-touching are not mutually exclusive.

We further adjusted our SVM set due to unexpected cases with face-touches. To avoid false-positives from occurring, we included the detection of hair-touches in order to

differentiate hair-touches from face-touches. Additionally, we discovered a case where face-touching and hand-touching were not mutually exclusive events (shown in Figure 5-6). As a result, we finalized our set of SVMs to the following:

Table 5.5: Final SVM set

SVM_1	SVM_2	SVM_3	SVM_4
Lean forward	Arms in lap	Open arms	Face touch
Lean back	Crossed arms	Hand touch	Hair touch

The following sections describe the training and testing of these 4 different SVMs for gesture recognition. We begin with Section 5.4.2 detailing the nature of the raw motion capture data and highlight some challenges and solutions to tracking failures in Section 5.4.3. We then continue with the training and testing of each SVM, which includes the following details:

- Set of features used for classification. Each of these features were filtered through a low-pass filter to reduce noise and scaled with a range of [-1,+1] to avoid features in greater numeric ranges dominating those with smaller ranges.
- Number of total participants, instances, and frames for each gesture in the testing and training sets.
- Training results using a leave-one-out (a whole participant) validation to determine the kernel parameters C and γ, choosing the parameters that give the best classification accuracy and also penalizing for over-fitting with high cost parameter C, which controls how rigid the hyperplane margins should be in allowing training error or misclassification.
- Testing results using 2 new participants that were not part of the training data.

- Results are reported with the following measures:
  - **a Confusion Matrix** with the columns of the matrix representing the instances of a predicted class and the rows representing the instances of an actual class
  - **% Recognition** which is the true positive rate (TP), or the proportion of positive cases that were correctly identified
  - **Score 1** which counts lost/invalid frames as missed detections in the scoring for the % Recognition
  - **Score 2** which does not count the lost/invalid frames in the scoring for the % Recognition
  - F_1 **Score**, or the F-measure, which scores accuracy in terms of both precision and recall (see Equation 5.1)

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{P * TP}{P + TP} \tag{5.1}$$

5.4.2 Motion-Capture Data

The features necessary for gesture recognition were derived from only the skeleton data from the kinect sensor since the gestures of interest only required tracking of points along the body instead of the face. As visualized in Figure 5-7 and organized in Table 5.6, we can track the real-world coordinates of 11 different points of the body and also know the orientations of a subset of 9 points. The 9 skeleton joints were tracked using OpenNI's NITE algorithm, and the additional 2 wrist points were tracked using HSV (hue, saturation, value) color-detection in finding the red and green wristbands.

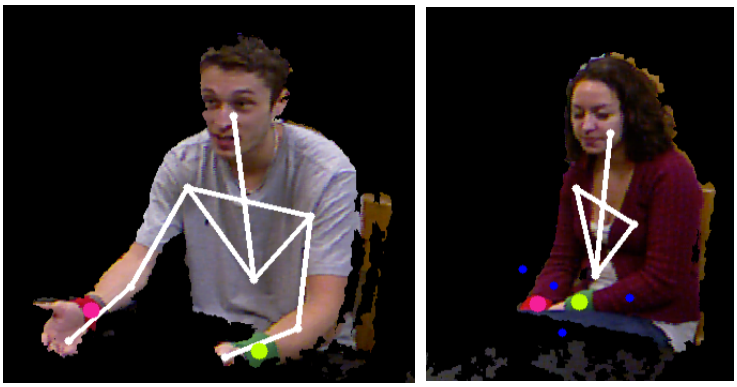


Figure 5-7: Nine joints tracked using OpenNI's skeleton-tracking along with an additional two wrist points using color-tracking. Points that are dark blue represent data of low confidence.

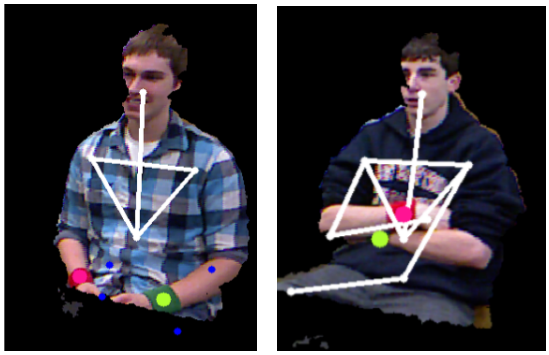
Table 5.6: List of tracked points and available data

Tracked part	3D Coordinates	Orientation
Head	yes	yes
Neck	yes	yes
Torso	yes	yes
Shoulder Left	yes	yes
Shoulder Right	yes	yes
Elbow Left	yes	yes
Elbow Right	yes	yes
Hand Left	yes	yes
Hand Right	yes	yes
Wrist Left	yes	no
Wrist Right	yes	no

The 3D (x, y, z) world coordinates (in millimeters) of a joint are relative to the location of the kinect sensor with +X pointing to the right of the sensor, +Y pointing up, and +Z pointing in the direction of increasing depth, or away from the sensor. The joint orientations are represented by a 3x3 rotation (orthonormal) matrix relative to the starting T-pose calibration (as shown in Figure 5-5) which is initialized as the identity matrix [OpenNI, 2011].

5.4.3 Tracking Challenges

As with any tracking system, there are cases in which the tracked points are completely lost or erroneous. Self-occlusions were a case that occurred quite frequently and would cause loss in tracking. Whenever the arms of an individual occluded the chest or were close enough to almost "merge" or be at the same depth level as the chest, the OpenNI tracking algorithm would lose or incorrectly track the locations of the elbows and hands as shown in Figure 5-8. As such, when detecting crossed-arms and arms-in-lap, we only used the head, shoulders, and torso data from skeleton-tracking in conjunction with the wristband locations from color-tracking. Unfortunately, we could not rely on this error state always corresponding to crossed-arms or arms-in-lap due to its inconsistency. Although very few, there were times when the skeleton was successfully tracked even with self-occlusions. As mentioned previously in Section 5.4.1, since this error state only applied to crossed-arms and arms-in-lap and not for open-arms, the set of features used for detection differed between those error state gestures and open-arms, resulting in a further split of SVMs.



(a) Arms in lap (b) Arms crossed

Figure 5-8: Examples of self-occlusions that caused errors in skeleton-tracking.

Environmental factors also posed some tracking problems, which in future studies can be easily mitigated. For a few participants, the backing of the chair in which

they sat on became a part of their chest. This would cause the center of mass of the participant to shift as to include the chair, making the locations of the shoulders and torso incorrect but not reported as such by the system. In future studies, a chair with a smaller backing can resolve this issue. Also, clothing like scarves caused poor skeleton-tracking, and participants that wore an article of clothing similar to that of the wristbands would have poor color-tracking. We asked participants to take off their jackets but unfortunately did not anticipant errors with scarves. In future studies, the wristbands should be more uniquely colored like hot pink or lime green. An example of each of these cases found with our participants are shown in Figure 5-9.

(a) Chair (b) Green shirt (c) Scarf

Figure 5-9: Environmental cases that caused errors with skeleton- and color-tracking.

Unfortunately, these error cases caused 4 participants' tracking to become very unreliable, and excluded from the training and testing corpuses. The remaining 49 participants used for training and testing had fair to moderate tracking performances. And even still, with every participant there were moments where the skeleton and wristband tracking systems would temporarily fail and report either low-confidence or invalid values. Such frames were excluded from the training data, and for the testing data, those frames were counted as missed detections.

5.4.4 SVM_1 Lean Detection

To detect a person leaning-forward, leaning-backward, or not leaning at all, we looked at the following two important features: (also illustrated in Figure 5-10):

a) the x-axis rotation of the torso joint to gain a sense of how much the body is leaning
b) the z-distance from head to torso to know how far the head is protruding from the body

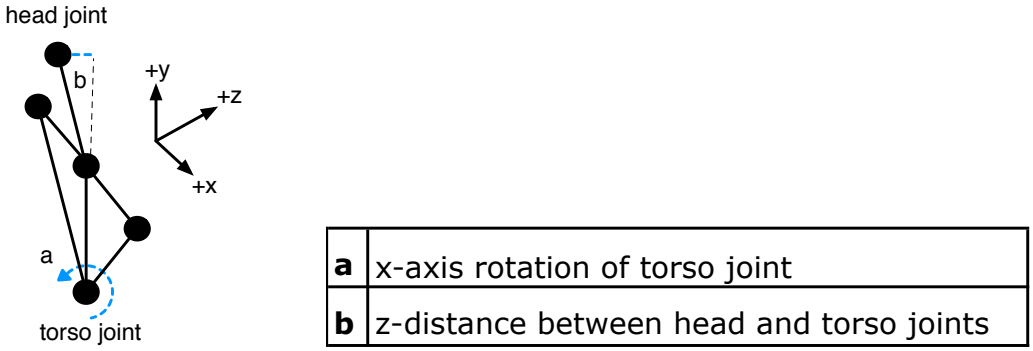


Figure 5-10: Features used for lean detection.

Our training set consisted of 20 participants with approximately 7,500 frames (5 minutes of data at 25 fps) per participant with each frame labeled as leaning-forward, leaning-back, or not-leaning. We had 39 instances of lean-forward, 17 of lean-back, and 51 of no-lean. It is important to note that these instances only tell how many different examples we have across participants and is not representative of how many training points or poses we have for each gesture which is dependent on the gesture's duration as well as whether the frame has missing data or invalid data. In total, we approximately have 56,000 frames of lean-forward, 68,000 frames of lean-back, and 70,000 frames of no-lean (see Table 5.7).

Table 5.7: Total number of instances and frames of the three leans from 20 participants used for training.

Lean	# instances	# frames
Fwd	39	56,000
Back	17	68,000
None	51	70,000

A leave-one-out (LOO) cross-validation was used as the training method to determine the SVM kernel parameter C. Unlike the other three SVMs, we decided to use a linear kernel function instead of radial for this SVM since intuitively one can imagine separating the leans with lines distinguishing where a particular lean begins and ends. We ranged the cost parameter C from $[2^{-5}, 2^{-3}, \ldots, 2^{15}]$, and each parameter was tested 20 times, where each time we would leave the training data from one participant out and train on the remaining 19. The omitted participant would then be used for testing to determine the parameter's ability to generalize to new data. Since the training set would contain a disproportionate amount of examples of the three different leans, we equalized the number of example frames coming from lean-forward, lean-back, and no-lean to the smallest common denominator before training, which according to Table 5.7 is about 56,000 example frames.

In a leave-one-out cross-validation with 20 subjects, we attained an average accuracy of 74.2% and an overall F_1 score of 74.0% in detecting the three different leans with cost function C = 0.1250 (see confusion matrix Table 5.8). We also performed another evaluation by testing two new participants that were not part of the training data. Collectively, the two participants had 5 instances of lean-forward (2,500 frames), 4 of lean-back (16,000 frames), and 8 of no-lean (800 frames). And we attained an average accuracy of 83.7% and an overall F_1 score of 69.0% (see Table 5.9). Since both the participants' data had zero frames of lost data, Score 1, which counts those frames as missed detections in the scoring, is the same as Score 2, which does not included those frames in the scoring.

Table 5.8: SVM_1 results with 20 subjects in a leave-one-out cross-validation

	No Lean	Lean Fwd	Lean back	% Recognition	F_1 Score
No Lean	45759	15960	9194	64.5%	64.5%
Lean Fwd	9929	45729	424	81.5%	77.2%
Lean Back	15372	688	52325	76.5%	80.3%
			Overall	74.2%	74.0%

Table 5.9: SVM_1 results with 2 new subjects

	No Lean	Lean Fwd	Lean Back	No Data	Score 1	Score 2	F_1 Score
No Lean	678	157	26	0	78.7%	78.7%	26.0%
Lean Fwd	37	2372	95	0	94.7%	94.7%	94.0%
Lean Back	3644	12	12682	0	77.6%	77.6%	87.0%
				Overall	83.7%	83.7%	69.0%

5.4.5 SVM_2 Arms-Pose Detection

As mentioned previously in the Tracking Challenges section 5.4.3, crossed-arms and arms-in-lap would cause the hands and the elbows to become erroneous in tracking due to self-occlusion. Therefore, we used the red and green wristbands as a means to track an individual's wrists to gather features for detecting the two arm poses. To detect crossed-arms and arms-in-lap, we used the following features (also illustrated in Figure 5-11):

a) distance between wristbands to capture how closely they are to each other

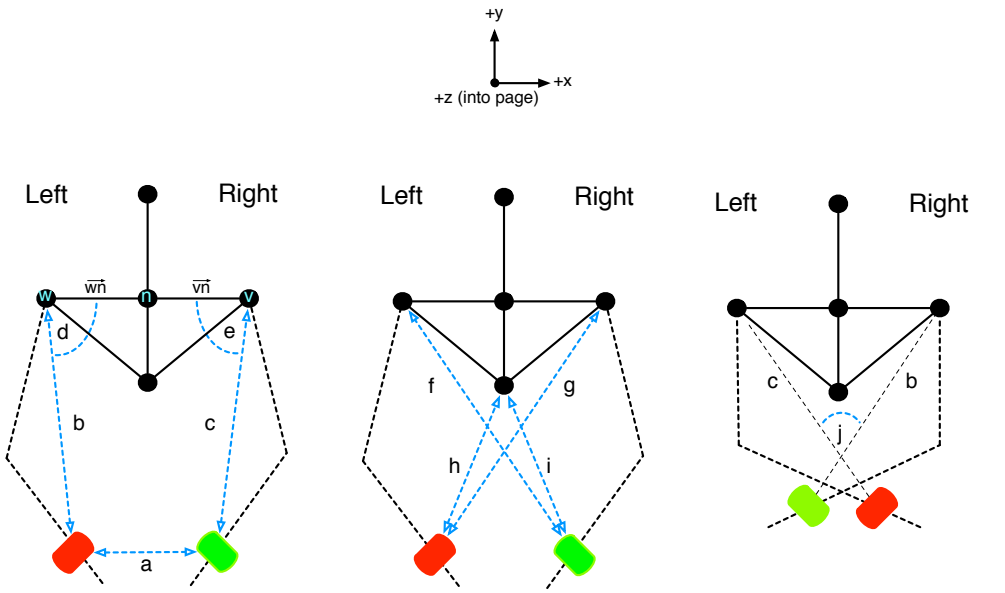
b/c) distance b/w the wrists and their respective shoulders to measure the arms' extension

d/e) angles the wrists make with the chest to know how tight or loose they're from the chest

f/g) distance between the wrists and their opposing shoulders (important for crossed-arms)

h/i) distance b/w the wrists and torso to gain a sense of how far they are from the body

j) angle between vectors b and c to know the angles made for a crossed-arm



a	distance b/w wristbands (wb)
b	distance b/w left wb and left shoulder
c	distance b/w right wb and right shoulder
d	angle b/w vectors b and wn
e	angle b/w vectors c and vn

f	distance b/w right wb and left shoulder
g	distance b/w left wb and right shoulder
h	distance b/w right wb and torso
i	distance b/w left wb and torso
j	angle b/w vectors b and c

Figure 5-11: Features used for arm-pose detection.

Features a, b, c, f, g, h, and i had to be further normalized per participant due to differences in body types and sizes. For example, feature a, which is the distance between the hands, can take on a much larger value with a larger participant with broad shoulders in comparison to a smaller participant with petite shoulders. As such, we normalized feature a by dividing it by the length of the participant's shoulders so that the values of feature a can carry the same meaning across participants. Features b, c, f, g, h, and i underwent similar normalization but with their respective normalization components.

Our training set consisted of 8 participants with approximately 7,500 frames per participant with each frame labeled as crossed-arms, arms-in-lap, or neither. We had 24 instances of crossed-arms, 79 of arms-in-lap, and 83 of neither. In total, we approximately have 6,000

frames of crossed-arms, 30,000 frames of arms-in-lap, and 44,000 frames of neither (see Table 5.10).

Table 5.10: Total number of instances and frames of the different arm-poses from 8 participants used for training.

Arms	# instances	# frames
Crossed	24	6,000
In Lap	79	30,000
None	83	44,000

A leave-one-out cross-validation was used as the training method to determine the SVM radial kernel parameters C and γ. We use a radial kernel function instead of linear as we expected the relationships between the features and class labels to be nonlinear. We ranged the cost parameter C from $[2^{-5}, 2^{-3}, \ldots, 2^{15}]$ and parameter γ from $[2^{-15}, 2^{-13}, \ldots, 2^{3}]$ and each parameter combination was tested 8 times, where each time we would leave the training data from one participant out and train on the remaining 7. The omitted participant would then be used for testing to determine the parameters' ability to generalize to new data. We again equalized the number of example poses to the smallest common denominator, which according to Table 5.10 is about 6,000 example frames.

In a leave-one-out cross-validation with 8 subjects, we attained an average accuracy of 58.7% and an overall F_1 score of 53.7% in detecting the different arm-poses with cost function C = 2 and $\gamma = 1.2207e^{-4}$ (see Table 5.11). In our evaluation with two new participants, we had 1 instance of crossed-arms (500 frames), 5 instances of arms-in-lap (1,800 frames), and 8 instances of neither (17,000 frames). And we obtained an overall F_1 score of 21.7% and an average accuracy of 19.4% and a small boost to 20.0% and when discounting the frames with missing data (see Table 5.12).

Table 5.11: SVM_2 results with 8 subjects in a leave-one-out cross-validation

	Neither	Crossed-Arms	Arms-in-lap	% Recognition	F_1 Score
Neither	28702	7972	7460	65.0%	70.8%
Crossed-Arms	541	3139	2297	52.5%	29.2%
Arms-in-lap	7698	4407	17192	58.7%	61.1%
			Overall	58.7%	53.7%

Table 5.12: SVM_2 results with 2 new subjects

	Neither	Crossed-Arms	Arms-in-lap	No data	Score 1	Score 2	F_1 Score
Neither	8722	5876	1761	633	51.3%	53.3%	63.7%
Crossed-Arms	581	42	0	0	6.7%	6.7%	1.3%
Arms-in-lap	1722	79	2	0	0.1%	0.1%	0.1%
				Overall	19.4%	20.0%	21.7%

5.4.6 SVM_3 Hand-touch and Open-arms Detection

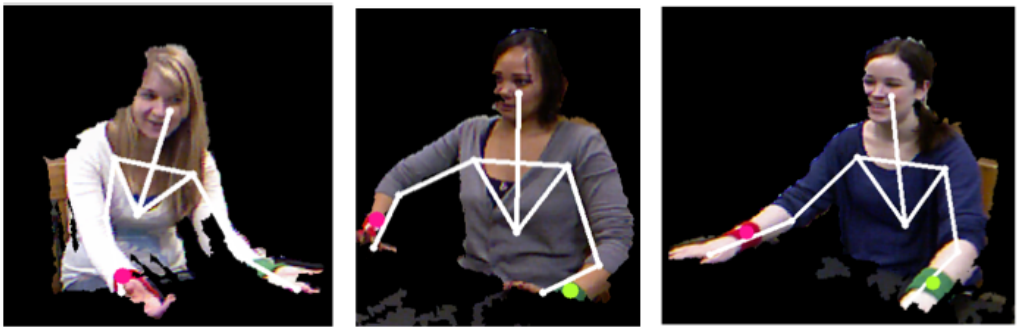


Figure 5-12: Examples of an open-arms pose.

An open-arms pose is described as having both arms outstretched in a way that fully exposes the chest as shown above in Figure 5-12. And in detecting an open-arms pose and hand-touches we used the following features (also illustrated in Figure 5-13):

a) distance between hands to capture how close they are to each other

b/c) distance between the hands and torso to gain a sense of how far they are from the body

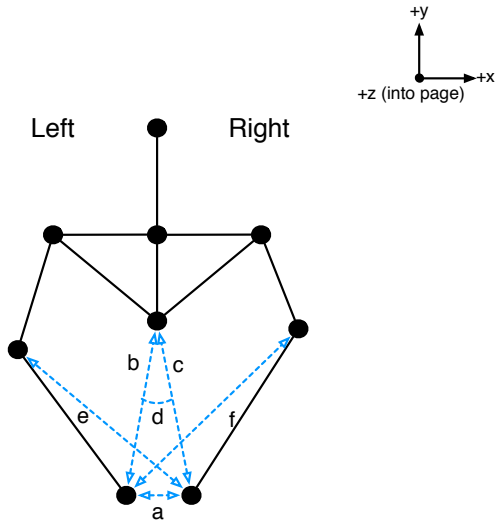
d) angle the hands make with the torso to gain a sense of how much they cover the body

e/f) distance between the hands and their opposing elbows to know how open the arms are

g/h) distance b/w the hands and their respective shoulders to measure the arms' extension

i/j) distance b/w the hands and their opposing shoulders to know how open the arms are
k/l) angles the hands make with the chest to know how tight or loose they're from the chest

And like with the features in SVM_2, we further normalized features a, b, c, e, f, g, h, i, and j in order to have values that remained consistent between participant.



a	distance b/w hands
b	distance b/w left hand and torso
c	distance b/w right hand and torso
d	angle b/w vectors b and c
e	distance b/w right hand and left elbow
f	distance b/w left hand and right elbow

g	distance b/w left hand and left shoulder
h	distance b/w right hand and right shoulder
i	distance b/w left hand and right shoulder
j	distance b/w right hand and left shoulder
k	angle b/w vectors g and wn
l	angle b/w vectors h and vn

Figure 5-13: Features chosen to detect open-arms and hand-touch.

Our training set consisted of 8 participants with approximately 7,500 frames per participant with each frame labeled as open-arms, hand-touch, or neither. We had 38 instances of open-arms, 261 of hand-touch, and 293 of neither. In total, we approximately have 2,000 frames of open-arms, 40,000 frames of hand-touch, and 33,000 frames of neither (see Table 5.13).

Table 5.13: Total number of instances and frames of the different poses from 8 participants used for training.

Pose	# instances	# frames
Open-arms	38	2,000
Hand-touch	261	40,000
None	293	33,000

Again, a leave-one-out cross-validation was used as the training method to determine the SVM kernel radial parameters C and γ. We ranged the cost parameter C from $[2^{-5}, 2^{-3}, \ldots, 2^{15}]$ and parameter γ from $[2^{-15}, 2^{-13}, \ldots, 2^{3}]$ and each parameter combination was tested 8 times, where each time we would leave the training data from one participant out and train on the remaining 7. The omitted participant would then be used for testing. We again equalized the number of examples to the smallest common denominator, which according to Table 5.13 is about 2,000 example frames.

In a leave-one-out cross-validation with 8 subjects, we attained an average accuracy of 62.5% and an overall F_1 score of 44.9% with cost function C = 32 and $\gamma = 0.0078$ (see Table 5.14). In our evaluation with two new participants, we had 13 instances of open-arms (1,500 frames), 57 of hand-touch (13,500 frames), and 74 of neither (5,000 frames). And we obtained an overall F_1 score of 43.0% and an average accuracy of 51.3% and a small boost to 52.3% when discounting the frames with missing data (see Table 5.15).

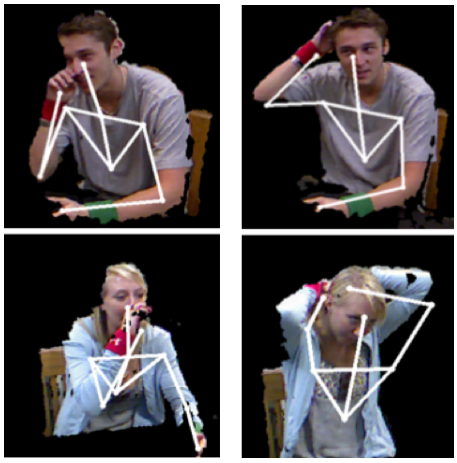
Table 5.14: SVM_3 results with 8 subjects in a leave-one-out cross validation

	Neither	Open-arms	Hand-touch	% Recognition	F_1 Score
Neither	11978	15400	5869	36.0%	50.3%
Open-arms	192	2015	103	87.2%	12.6%
Hand-touch	2184	12352	26029	64.2%	71.7%
			Overall	62.5%	44.9%

Table 5.15: SVM_3 results with 2 new subjects

	Neither	Open-arms	Hand-touch	No data	Score 1	Score 2	F_1 Score
Neither	1895	2298	728	361	35.9%	38.5%	47.7%
Open-arms	191	1097	220	0	72.7%	72.7%	20.3%
Hand-touch	946	5912	6108	516	45.3%	45.6%	61.0%
				Overall	51.3%	52.3%	43.0%

5.4.7 SVM_4 Touch Detection

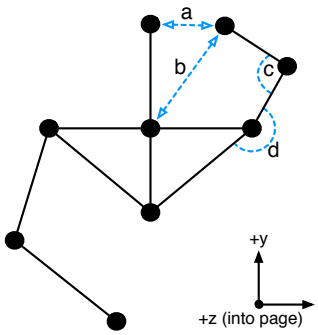


(a) Face touch (b) Hair touch

Figure 5-14: Examples of how hair and face touches can differ when considering the shoulder and elbow angles.

In order to mitigate the number of false-positives for face-touch, we detected for both hair-touching and face-touching events. Unfortunately, since the hair is so close to the face, some instances will require more information, beyond just pose, to know whether a person is touching their face or their hair. However, there are some instances where we can differentiate the two from pose-information alone. As shown in Figure 5-14, when individuals scratch or play with their hair they sometimes lift their elbow above their shoulders, but when they are touching their face, they keep their elbow close to the table and reach up with just their forearm. Thus, to differentiate these two cases, we looked at the following features (also illustrated in Figure 5-15):

a/b) distance b/w the hand and the head/neck joints to know how close the hand is to face
c/d) shoulder and elbow angles to help differentiate between face and hair touches



a	distance b/w hand and head
b	distance b/w hand and neck
c	elbow angle
d	shoulder angle

Figure 5-15: Features chosen to detect face-touch and hair-touch.

Our training set consisted of 14 participants with approximately 7,500 frames per participant with each frame labeled as face-touch, hair-touch, or neither. We had 90 instances of face-touch, 47 of hair-touch, and 161 of neither. In total, we approximately have 3,000 frames of face-touch, 1,000 frames of hair-touch, and 170,000 frames of neither (see Table 5.16).

Table 5.16: Total number of instances and frames of the touch poses from 14 participants used for training.

Touch	# instances	# frames
Face	90	3,000
Hair	47	1,000
None	161	170,000

Again, a leave-one-out cross-validation was used as the training method to determine the SVM kernel radial parameters C and γ. We ranged the cost parameter C from

$[2^{-5}, 2^{-3}, \ldots, 2^{15}]$ and parameter γ from $[2^{-15}, 2^{-13}, \ldots, 2^{3}]$ and each parameter combination was tested 14 times, where each time we would leave the training data from one participant out and train on the remaining 13. The omitted participant would then be used for testing. We again equalized for same-sized example sets, which according to Table 5.16 is about 1,000 example frames.

In a leave-one-out cross-validation with 14 subjects, we attained an average accuracy of 81.6% and an overall F_1 score of 50.2% with cost function C = 8 and $\gamma = 0.1250$ (see Table 5.17). In our evaluation with two new participants, we had 9 instances of face-touch (300 frames), 7 of hair-touch (270 frames), and 19 of neither (38,000 frames). And we obtained an overall F_1 score of 50.4% and an average accuracy of 51.5% and a boost to 60.5% when discounting the frames with missing data (see Table 5.18).

Table 5.17: SVM_4 results with 14 subjects in a leave-one-out cross-validation

	Neither	Face-touch	Hair-touch	% Recognition	F_1 Score
Neither	140267	28198	1309	82.6%	90.5%
Face-touch	30	2459	271	89.1%	14.6%
Hair-touch	10	275	774	73.1%	45.4%
			Overall	81.6%	50.2%

Table 5.18: SVM_4 results with 2 new subjects

	Neither	Face-touch	Hair-touch	No Data	Score 1	Score 2	F_1 Score
Neither	35279	383	23	2287	92.90%	98.4%	99.4%
Face-touch	1	162	71	71	53.1%	69.2%	35.4%
Hair-touch	5	137	23	106	8.5%	13.9%	16.3%
				Overall	51.5%	60.5%	50.4%

5.4.8 Discussion

Out of the 4 SVMs for gesture detection, SVM_1 for lean detection performed the most reliability with an overall F_1 score of 69.0% and an average recognition accuracy of 83.7% in differentiating leaning-forward, leaning-back, and no-leaning with two new participants. One of the major challenges in creating a generic lean-detector is that leans can be a very relative description. We found the video-coding for lean-forward, lean-back, and no-lean to be very dependent on the participant's natural posture; that is an individual's no-leaning pose that is a little hunched over would be considered a lean-forward for another participant who had better posture. So then, a coder would adjust their sense of leans to a particular participant, making lean definitions vary among participants. As such, subjects with such confusing margins in knowing when they are leaning-forward or not-leaning at all or when they are leaning-back, will experience the least reliable detection. And as seen in the results of the confusion matrix in Table 5.8, the lean-detection is very good at differentiating lean-forward from lean-back and vice-versa but encounter some difficulties in creating a good boundary for between lean-forward and no-lean as well as lean-back and no-lean.

The SVMs that performed the least reliably are SVM_2 for crossed-arms and arms-in-lap detection and SVM_3 for open-arm and hand-touch detection with average accuracies of 20.0% and 52.3% respectfully and with overall F_1 score of 21.7% and 43.0% respectfully. Unfortunately, there were sparse examples of crossed-arms, arms-in-lap, and open-arms. While there were many instances of these gestures within a single participant, we did not have many examples across participants, hindering the SVM from creating a generalized model for these gestures. For future-work, since crossed-arms and arms-in-lap occur too infrequently in a natural setting but are gestures that can be easily generated by actors, we can increase the number of training examples by including additional actor-generated ones. Open-arms unfortunately are difficult to naturally generate on-the-fly and are also the rarest to observe in the "wild" and

also suffer from a somewhat ambiguous definition (with disputes amongst coders). However, what this pose tries to capture is the "openness" of a person's stance. Is a person closing themselves off by crossing their arms as if using their arms as a barrier and shield? Or are they comfortably exposing their chest almost vulnerable to a frontal attack? Thus, when trying to capture such stances, a more appropriate means, beyond just looking at arm-poses, would be to see how much area their poses take up. If a person is tightly crossing their arms and hunching in, their total area would be much smaller than someone with a wide ready-to-give-a-hug pose. By segmenting an individual's body from the scene, we can measure how the total area of an individual's body expands or contracts throughout an interaction to gain a more accurate gauge in which the coded open-arm gesture attempted to measure.

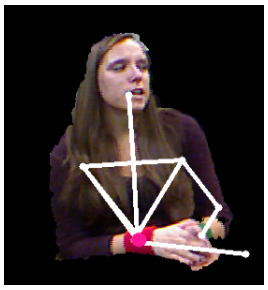


Figure 5-16: Example of a participant touching their hands together but the skeleton-tracking incorrectly labeling the location of the left hand.

In addition to this separate form of open-detection (that is beyond skeleton-tracking), we also suggest a specific hand-tracking algorithm to find the exact locations of the hands. Although we had many examples of hands-touching, there were many instances in which the skeleton-tracking would incorrectly label the locations of the hands. As shown in Figure 5-16, the erroneously-tracked hand locations would lie arbitrarily along the forearm instead of the true spot, making the distance between these pseudo-hands much larger, which was an important feature for hand-touch detection. As such, for future-work, we suggest using a separate tracking method that

is specifically for hands in order to have better reliability for hand-touch detection.

Lastly, SVM_4 can reliably detect when a person is neither touching their face nor hair with an average recognition accuracy of 98.4% (F_1 score of 99.4%) but cannot accurately differentiate face-touches from hair-touches by means of pose-information alone. For future-work, we suggest to segment the hair from the skin by means of surface textures to establish the boundaries of the two regions and by doing so gain a more accurate idea of what the hand is touching.

Chapter 6

Validation of System

6.1 Overview

In this section, we want to test how accurately our entire system, with low-level gesture recognition for high-level trust recognition, can autonomously predict whether an individual finds another to be trustworthy or untrustworthy. By testing the combined system with new participants that were not a part of the training corpuses, we can validate how well the combined system can automate the detection of trust as well as generalize to new situations and new people. Unfortunately, due to the low-recognition reliability reported in Section 5.4 as well as unexpected behaviors introduced in Study 2 (which is thoroughly discussed below), we were unable to perform a full system test. However, we describe the reasons that attribute to the system's inability to generalize to the participants in Study 2 and give suggestions for future-studies that can either easily by-pass the problems we faced or more elegantly overcome them in creating a more robust Trust model. With our future suggestions, we believe that it is possible to have a working automated Trust recognition system.

6.2 Trust Model Validation

In Section 4.1, we trained and tested through cross-validation a hidden markov model to determine whether an individual will have high or low trust judgements of their novel partner by observing how the 7 predictive gestural cues unfolded in the social interaction. Using coded observations from Study 1, we were able to detect with 94% accuracy whether a participant would give 2 or 4 tokens (indicative of low or high trust judgements) to their fellow participant in the Give Some Game. With completely new participants obtained in Study 2, we tested how well the Trust model could predict whether a participant would give 2 or 4 tokens.

6.2.1 Results

Out of the 41 remaining participants reported in Section 5.3.4, we had a total of 11 participants that gave two tokens to their partner and 4 participants that gave away four. And each test case consisted of a sequence of observed (human-coded) gestural cues that the participant emitted for the entire 5-minute interaction. Out of the 11 participants that gave away only two tokens, the HMM model correctly predicted 8 of them would give away two tokens. And out of the 4 participants that gave four tokens, the Trust model incorrectly predicted that all these participants would give only two tokens (see confusion matrix Table 6.1). We thoroughly discuss below how the differences in the setup of Study 1 and Study 2 generated significant changes in behaviors between the participant sets, which in turn caused the Trust model to perform with such poor accuracy.

Table 6.1: Confusion matrix for testing the HMM Trust model on 15 new participants.

		Predicted		
		Low-Trust	High-Trust	% Recognition
Actual	Low-Trust	8	3	72.7%
	High-Trust	4	0	0%

6.2.2 Discussion

When comparing Study 1 and Study 2, there were three minor differences which at the time we thought were inconsequential, but after carefully reviewing the videos of participants in Study 1 versus the ones in Study 2, we found major differences in their starting poses as well as their hand-touching behaviors. These three minor differences were:

1) the removal of the sheet of paper with conversation topics
2) the introduction of wristbands for participants to wear
3) the introduction of the motion capture technology.

We were most concerned with item 3 since the kinects and swissrangers were very visible and pointed towards the participants, which could potentially cause participants to become more aware of their movements preventing natural gestures from occurring. But in our follow-up interview we found that majority of participants were not bothered by the presence of the cameras since the equipment was off to the side, and the participants were more focused on interacting with their partner.

Figure 6-1: 16% more of participants in Study 2 assumed this pose at the start of the interaction in comparison to Study 1.

Surprisingly, the introduction of the wristbands and the removal of the sheet of paper caused the most disruption in the participant's behaviors. As mentioned in Study 2's protocol in Section 5.3.3, after the participants sat down in their chairs, we asked them to wear the wristbands that were on the table. And after doing so, we found that participants then left their hands resting on the table significantly more than participants in Study 1. Out of all the 56 participants in Study 2, 44 (or 79%) of participants began with their hands on the table, while (by randomly choosing an equal number of participants) only 35 (or 63%) of participants began with their hands resting on the table in Study 1. By having the participants wear the wristbands at the table, we biased them in starting the conversation in the particular pose of resting their arms on the table with their hands-touching as show in Figure 6-1. This pose

bias has implications across all the predictive gestures as it affects the starting leaning pose, arm pose, and touching behavior.

Figure 6-2: Participants in Study 1 leaving their hands at the sides of the piece of paper.

The differences in the hand-touching behavior between the two studies were further exacerbated by the removal of the sheet of paper with the conversation topics. The sheet of paper acted as barrier between the two hands for participants in Study 1. Instead of fiddling with their hands, participants in Study 1 would fiddle with the sheet of paper. And even after fiddling with it, they would leave their hands on the sides of the paper as shown in Figure 6-2.

And in removing this "barrier" in the second study, we found, with an unpaired 2-tailed t-test of unequal variance, that participants in Study 2 touched their hands significantly more than participants in Study 1 [$p < 0.0013$, $\mu_{study1} = 7.40$, $\mu_{study2} = 11.42$] (see Figure 6-3). Since hand-touching is considered to be a negative cue, this increase of hand-touch events caused the Trust model to incorrectly predict that all trusting participants did not trust their partners. The significance of hand-touching as a negative cue in the presence of the paper "barrier" in Study 1 was lost in Study 2.

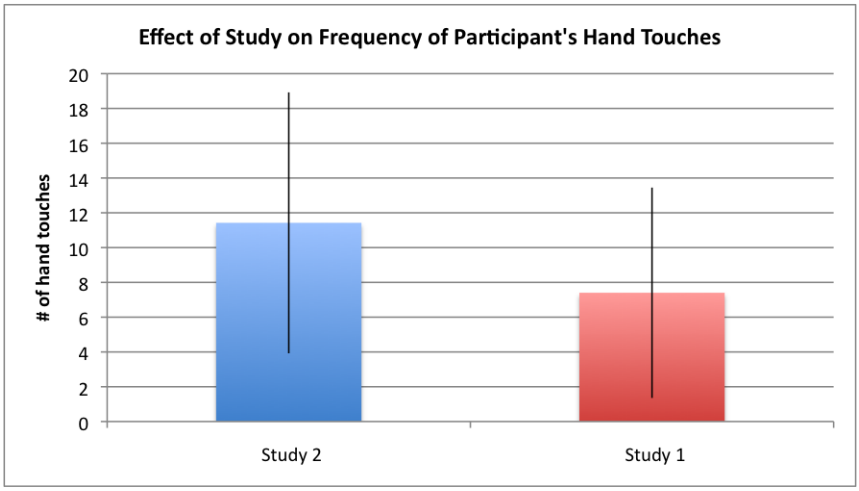


Figure 6-3: Participants in Study 2 touched their hands significantly more than those in Study 1

For future-work, there are a few ways to resolve these issues. The easiest fix is to have the participants wear the wristbands before they sit down at the table and also reintroduce the sheet of paper to regain the significance of hand-touching. A more elegant solution is to combine the behaviors seen in Study 1 and Study 2 into a new and better generalized model that is not solely based on the emission of a gesture, but also based on the duration as well. Currently, the model as it is now cannot capture the significance of a participant crossing their arms the entire time of the interaction. The model relies on the fact that the participant would perform another gesture and then would reassume the crossed-armed pose, and the stickiness or desire to go back to that pose would signal distrust. Although the frequency and the pattern of emitted cues is certainly a revealing tale of a person's level of trust as discussed in Section 4.4, the Trust model in its detection could be more robust with additional information of duration. And unfortunately, since the current Trust model could not generalize to the new study due to changes in hand-touch behaviors, we could not perform a whole system test as the high-level trust recognition would be incorrect no matter how well or poorly the low-level gesture recognition system performed.

Chapter 7

Conclusion

7.1 Contributions

Our work has taken initial strides towards understanding and modeling nonverbal behavior that is predictive of interpersonal trust. To the best of our knowledge, this work takes the first attempt in developing a computational model for recognizing interpersonal trust in social interactions. We began by leverage pre-existing datasets to understand the relationship between synchronous movement, mimicry, and gestural cues with trust. We found that although synchronous movement was not predictive of trust,

synchronous movement was positively correlated with mimicry.

That is, people who mimicked each other more frequently also move more synchronously in time together. Also, same-gender dyads synchronize more often than mixed-gender dyads, revealing the

bidirectional relationship between synchrony and feelings of connectivity.

Prior work has shown conscious synchronization, like rocking in chairs together, can be used as a means to prime individuals to have increase feelings of connectivity towards another. And our work suggests that an enhanced sense of connectivity, established through similarity in gender, can in turn cause more unconscious synchronization between dyads. We also introduced a new paradigm in robustly

measuring synchrony between dyads through background subtraction.

By using this computer vision algorithm, we can measure synchronized movements between dyads more reliably than the typical human-rating paradigm.

In investigating the relationship between mimicry and trust, we found that although mimicry was not predictive of trust in our study,

> *mimicry is predictive of liking.*

Participants that mimicked their partner more, also rated their partner as more likable, and so did the partner in return. Revealing the versatile nature of unconscious mimicry, our study situation, which was not conducive to active trust building and had a more socially-friendly nature, caused participant to utilize mimicry as a mechanism to convey liking instead of trust.

We reconfirmed DeSteno et al. [2011]'s four negative gestural cues that when taken together are predictive of lower levels of trust:

> *leaning backwards, face touching, hand touching, and crossing arms.*

And found three positive cues,

> *leaning forward, arms in lap, and open arms*

that when taken together are predictive of higher levels of trust. And by observing how these 7 important gestures unfold throughout the duration of a social interaction, we are able to predict whether an individual will trust their novel partner or not in behaving cooperatively in an economic "Give Some Game." Through probabilistic graphical modeling, we developed a

> *hidden markov model capable of differentiating with 94% accuracy (with chance at 68%) whether an individual will give 2 or 4 tokens (indicative of low or high trust judgements) to their novel partner by observing the sequence of predictive cues he/she emits.*

By comparing the resulting $HMM_{lowtrust}$ and $HMM_{hightrust}$ models, we found that not only does the frequency in the emission of the predictive gesture matter, but also

the sequence in which we emit negative to positive cues matter

The resulting model revealed that there is a more alternating pattern of positive to negative cues being emitted by a trusting individual, while a back-to-back succession of negative cues being displayed when an individual does not trust their partner.

In hopes of automating this detection of trustworthiness, we ran a new study in which we captured participants naturally performing these 7 predicative gestures in a dyadic social interaction by using the Microsoft kinect for skeleton-tracking in conjunction with wristband color-tracking. By training multiple support vector machines (SVMs), we were able to reliably detect when an individual

leans-forward, leans-back, and has no-lean with an average recognition accuracy of 83.7% and an overall F_1 score of 69.0%

and unreliably detect the remaining predictive gesture. However, we gave suggestions for improvements that can potentially make the recognition of all 7 predictive gestures more accurate for future-work.

Unfortunately, we were unable to perform the final testing of the entire system, with low-level gesture recognition for high-level trust recognition, due to the introduction of new factors in Study 2 that artificially increased the frequency of the negative hand-touching gesture. Our Trust model, which is based on the frequency and the pattern of emitted cues, could not accurately predict whether a participant in Study 2 found their partner to be trustworthy or not. In the section below, we propose and discuss the potential of a more robust Trust model for future-work.

Although we were only able to reliably recognize 2 out of the 7 predictive gestures, we are encouraged by the fact that those successfully modeled gestures had the most training examples, and we expect that with more examples as well as some improved

tracking techniques, we can eventually accurately recognize all the gestures of import. And by remodeling the Trust model to include information about a gesture's frequency as well as its duration, we hope to eventually show a working automated system, with low-level gesture recognition for high-level trust recognition, that is capable of predicting whether an individual finds another to be a trustworthy or untrustworthy partner.

7.2 Concluding Remarks

Much like a personal computer, robots can become a technology that is a part of our daily lives. And as robots began to interact with us, they will need to be capable of perceiving and understanding our social nuances to be functional agents in our talkative and expressive world. And with much of our communication beyond spoken literal words, which occupy only a small slice of our entire communication bandwidth, social robots need to be appropriately designed to be capable of understanding and responding both verbal and nonverbally as they collaborate with us as trusted partners. By modeling the dynamics of nonverbal behavior between people and how it can impact the quality of a social interaction, we not only better understand the complexities of our own social communication, but we can in turn design for more effective human-robot interactions.

7.3 Future Work

Synchronous movement: Recent interest has grown in understanding the role of synchronous movement in social interactions [Michalowski et al., 2007; Valdesolo et al., 2010]. For future-work, we would like to further investigate whether certain types of synchronized movement can be predictive of trust. Our study measured the overall synchrony of dyads, but it could be the case that the sync score included meaningless synchronized movements that are not related to trust (i.e. both participants shaking their foot). We would like to better understand the different types of synchronous movement between interacting individuals and how that can impact the quality of the interaction.

Also to further confirm the bidirectional relationship of synchrony and connectivity, we would like to run a two-condition study with dyads of either similar or dissimilar personalities which we expect to result with similar-typed personalities synchronizing more in their movements.

Mimicry: Although our study setup caused participants to use mimicry as a mechanism in conveying liking instead of trust, we are still interested in understanding mimicry's role in trust. In particular, we would like to discover what types of unconscious mimicry could be predictive of cooperative or uncooperative behaviors. There is strong evidence in past literature that mimicry is indeed related to trust, but most findings utilize passive simple behaviors like foot-shaking or face-rubbing as mimicry indicators. But this does not reveal how meaningful mimicked gestures beyond mere foot-shaking are indicative of higher or lower levels of trust.

Gesture recognition: One major challenge with gesture recognition was the complete loss or errors in skeleton-tracking in cases of self-occlusions. Although the wristbands mitigated some of these failure cases, by having the entire skeleton suc-

cessfully tracked, we could have had some important additional features that would have assisted in obtaining more reliable recognition rates. At the time of Study 2, OpenNI was currently the only available skeleton-tracking algorithm for the kinect, but recently Microsoft has released their software for skeleton-tracking. Along with the improvements suggested in Section 5.4.8 and trying out Microsoft's SDK, we hope that for future-work we can obtain better recognition of the 7 predictive gestures.

Trust model: As mentioned previously, a more generalized Trust model could be obtained by not only observing the sequence of predictive gestures, but also the duration in which an individual holds these gestures. The current model as it is now cannot capture the significance of a participant crossing their arms the entire time of the interaction. The model relies on the fact that the participant would perform another gesture and then would reassume the crossed-armed pose, and the stickiness or desire to go back to the pose would signal distrust. Although frequency and the pattern of emitted cues is certainly a revealing tale of a person's level of trust, the Trust model in its detection of trust could be more robust with this additional information of duration.

Appendix A

Supplement Material

A.1 Data Measurements

Dyad #	Sync Score	Mimic Score
18	0	20
73	0	27
55	1	30
57	1	24
6	2	45
9	4	23
69	6	33
4	10	24
43	11	31
8	20	18
61	21	55
66	23	52
54	39	21
52	50	11
51	58	50
2	62	19
1	67	64
60	70	19
75	93	60
17	107	26
13	121	59
15	122	51
71	139	39
56	162	56
64	172	62
7	172	48
44	189	35
59	191	53
58	197	27

Figure A-1: Final mimicry score and synchronous movement score for each dyad.

A.2 Trust HMM Model Parameters

Transition Matrix A		
0.9482	0.0202	0.0315
0.1717	0.0759	0.7524
0.1735	0.5824	0.2441

Priors Pi
0.0473
0.9456
0.0072

Observation Matrix B					
0.0028	0.2720	0.0183	0.0107	0.0617	0.2511
0.5690	0.3289	0.0465	0.0019	0.0000	0.0514
0.0343	0.1791	0.0005	0.3454	0.0004	0.0545

Figure A-2: HMM model $\lambda = (A, B, \pi)$ for high trust $N_H = 3$

Transition Matrix A						
0.0312	0.4255	0.3498	0.0792	0.0866	0.0006	0.0272
0.4111	0.0057	0.2825	0.0435	0.1765	0.0489	0.0319
0.3136	0.1517	0.2059	0.0047	0.1519	0.1640	0.0082
0.0514	0.0402	0.0096	0.2795	0.0762	0.0552	0.4880
0.3774	0.1781	0.1924	0.0108	0.1099	0.0214	0.1101
0.2063	0.1016	0.2696	0.1795	0.0962	0.0006	0.1463
0.1074	0.0178	0.0054	0.2907	0.0437	0.0189	0.5160

Priors Pi
0.0256
0.0087
0.0197
0.0208
0.0780
0.8402
0.0071

Observation Matrix B						
0.0159	0.5054	0.0009	0.0118	0.0009	0.4420	0.0231
0.1122	0.0323	0.0258	0.0039	0.0073	0.0055	0.8131
0.0091	0.2886	0.0344	0.1686	0.0028	0.0912	0.4053
0.0849	0.0171	0.0000	0.0022	0.2285	0.4519	0.2154
0.1518	0.2898	0.0112	0.0615	0.0048	0.1832	0.2976
0.5650	0.4142	0.0001	0.0041	0.0006	0.0053	0.0107
0.0804	0.0179	0.0000	0.0151	0.2887	0.4593	0.1386

Figure A-3: HMM model $\lambda = (A, B, \pi)$ for low trust $N_H = 7$

Bibliography

Bailenson, J. and Yee, N. (2005). Digital chameleons. *Psychological Science*, 16(10):814.

Bailenson, J. N., Beall, A. C., Loomis, J., Blascovich, J., and Turk, M. (2004). Transformed Social Interaction: Decoupling Representation from Behavior and Form in Collaborative Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 13(4):428–441.

Benton, S. (2008). Background subtration, part 1: Matlab models. *EE Times Design*.

Bernieri, F. J. (1988). Coordinated movement and rapport in teacher-student interactions. *Journal of Nonverbal Behavior*, 9(2):113–138.

Chartrand, T., Maddux, W., and Lakin, J. (2005). Beyond the perception-behavior link: The ubiquitous utility and motivational moderators of nonconscious mimicry. *The new unconscious*, pages 334–361.

Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: the perception-behavior link and social interaction. Technical Report 6, Department of Psychology, New York University, New York 10003, USA.

DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D., Baumann, J., Dickens, L., and Lee, J. J. (2011). Detecting the Trustworthiness of Novel Partners in Economic Exchange. Submitted for publication.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern Classification*, volume November. Wiley.

Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R., and Morency, L. (2006). Virtual rapport. In *Intelligent Virtual Agents*, pages 14–27. Springer.

Hsu, C., Chang, C., Lin, C., and Others (2003). A practical guide to support vector classification.

Kanda, T., Hirano, T., Eaton, D., and Ishiguro, H. (2004). Interactive robots as social partners and peer tutors for children: A field trial. *Human-Computer Interaction*, 19(1):61–84.

LaFrance, M. (1979). Nonverbal Synchrony and Rapport: Analysis by the Cross-Lag Panel Technique. *Social Psychology Quarterly*, 42(1):66–70.

Lakin, J., Jefferis, V., Cheng, C., and Chartrand, T. (2003). The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *Journal of nonverbal behavior*, 27(3):145–162.

Lakin, J. L. and Chartrand, T. L. (2003). Using Nonconscious Behavioral Mimicry to Create Affiliation and Rapport. *Psychological Science*, 14(4):334–339.

Lee, S. (2007). The relations between the studentâĂŞteacher trust relationship and school success in the case of Korean middle schools. *Educational Studies*, 33(2):209–216.

Maddux, W., Mullen, E., and Galinsky, A. (2007). Chameleons bake bigger pies and take bigger pieces: Strategic behavioral mimicry facilitates negotiation outcomes. *Journal of Experimental Social Psychology*, 44(2):461–468.

MESA (2011). Mesa imaging. `http://www.mesa-imaging.ch/index.php`.

Michalowski, M. P., Sabanovic, S., and Kozima, H. (2007). A dancing robot for rhythmic social interaction. *Proceeding of the ACMIEEE international conference on Humanrobot interaction HRI 07*, page 89.

Microsoft (2011). Kinect for windows sdk beta. `http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk/default.aspx`.

Muñoz, J. H. (2011). Integrated Vision Framework for a Robotics Research and Development Platform. Masters Thesis.

Murphy, K. (2007). Bayes net toolbox for matlab. `http://code.google.com/p/bnt/`.

Noldus (2011). Noldus information technology. `http://www.noldus.com/`.

OpenNI (2011). Open natural interaction. `http://www.openni.org/`.

Pan, W., Dong, W., Cebrian, M., Kim, T., and Pentland, A. (2011). Modeling Dynamical Influence in Human Interaction. pages 1–20.

Pentland, A. S. (2008). *Honest Signals*. MIT Press.

Pinyol, I. and Sabater-Mir, J. (2011). Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*.

PointGrey (2011). Firefly mv cmos camera. `http://www.ptgrey.com/products/fireflymv/`.

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*.

Sen, S. and Sajja, N. (2002). Robustness of reputation-based trust: Boolean case. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 288–293. ACM.

Soman, V. (2009). Social Signaling: Predicting the Outcome of Job Interview from Vocal Tone and Prosody. *Electrical Engineering*, (March).

Tickle-Degnen, L. (1990). The Nature of Rapport and Its Nonverbal Correlates. *Psychological Inquiry*, 27(4):316–293.

Trout, D. L. (1980). The effect of postural lean and body congruence on the judgment of psychotherapeutic rapport. *Journal of Nonverbal Behavior*, 41(3):700–190.

Valdesolo, P. and DeSteno, D. (2011). Synchrony and the social tuning of compassion. *Emotion (Washington, D.C.)*, 11(2):262–6.

Valdesolo, P., Ouyang, J., and DeSteno, D. (2010). The rhythm of joint action: Synchrony promotes cooperative ability. *Journal of Experimental Social Psychology*, 46(4):693–695.

Van Lange, P. A. M. and Kuhlman, D. M. (1994). Social value orientations and impressions of partner's honesty and intelligence: A test of the might versus morality effect. *Journal of Personality and Social Psychology*, 67(1):126–141.

Wilkes, D. M., Franklin, S., Erdemir, E., Gordon, S., Strain, S., Miller, K., and Kawamura, K. (2010). Heterogeneous Artificial Agents for Triage Nurse Assistance. *2010 10th IEEERAS International Conference on Humanoid Robots*, pages 130–137.

Wiltermuth, S. S. and Heath, C. (2009). Synchrony and cooperation. *Psychological Science*, 20(1):1–5.