THESIS

AN EMPATHIC AVATAR IN TASK-DRIVEN HUMAN-COMPUTER INTERACTION

Submitted by

Heting Wang

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2020

Master's Committee:

Advisor: Ross Beveridge
Co-Advisor: Francisco Ortega

Julia Sharp
Christopher Peterson

ABSTRACT

AN EMPATHIC AVATAR IN TASK-DRIVEN HUMAN-COMPUTER INTERACTION

In Human-Computer Interaction, it is difficult to give machines emotional intelligence to resemble human affects, such as the ability of empathy. This thesis presents our work of an emotionally expressive avatar named Diana that can recognize human affects, and show her empathy by using dynamic facial expressions. Diana's behaviors and facial expressions were modeled from Human-Human multimodal interactions to help to provide human-like perceptions in users. Specifically, we designed her empathic facial expressions as a linear combination of the action units in the Facial Action Coding System [1], with the action units that were previously found to improve the accuracy and judgments of human likeness.

Our work studies the role of affect between a human and Diana working together in a blocks world. We first conducted an elicitation study to extract naturally occurring gestures from naive human pairs. The pair of human collaborated on a task remotely through video communication to build wooden blocks. The video footage of their interactions composed a dataset named EGGNOG [2]. We provided descriptive and statistical analysis of the affective metrics between human *signalers* and *builders* in EGGNOG. The metrics included measures of valence (positive or negative experience) and intensities of 7 basic emotions (joy, fear, disgust, anger, surprise, and contempt). We found: 1) Overall the *signalers* had a broader range of valence and showed more varied emotions than the *builders*. 2) The intensity of *signalers'* joy was greater than that in *builders*, indicating a happier *signaler* than a *builder*. 3) For individuals, the person was happier to act as a *signaler* in a task than act as a *builder*. Additionally, valence was more associated with a person's role in a task and less associated with personality traits. Other emotions were all weak and no significant difference was found between *signalers* and *builders*.

To adapt to the user's affects in the later Human-Avatar interaction, we modeled Diana's empathic behaviors based upon findings in EGGNOG and the Appraisal theory [3]. We created a Demo mode of Diana whose affective states, i.e., facial expressions that simulated empathy, dynamically transitioned between 5 finite states (neutral, joy, sympathy, concentration, and confusion) with respect to the user's affects and gestures. We also created a Mimicry mode of Diana who mimicked the user's instant facial expressions. Human subject studies involving three modes of this avatar (Demo, Mimicry, and Emotionless) were conducted with 21 participants. The difference in votes from a 5-point Likert scale perception questionnaire or a NASA TLX perceived load survey was both statistically insignificant. However, compared to the Mimicry Diana and the Emotionless Diana, a descriptive analysis indicated users spent more time engaging with the empathic Diana, and both the Demo and Mimicry mode of Diana were preferred by users over the Emotionless Diana. Some participants commented about Diana's facial expressions as natural and friendly while 3 other participants were elicited uncomfortable feelings and mentioned the Uncanny Valley effect. Results indicated our approach of adding affects to Diana was perceived differently by different people and received both positive and negative feedback.

Our work provided another implementable direction of the human-centered user interfaces with complex affective states. However, there was no evidence that the empathic facial expressions were more preferred by participants than the mimicked facial expressions. In the future, Diana's empathic facial expressions may be refined by modeling more human-like action unit movements with the help of deep learning networks, and the user perception in subjective reports may get improved.

ACKNOWLEDGEMENTS

I would like to deeply appreciate the supports form Dr. Ross Beveridge, Dr. Francisco Ortega, Dr. Julia Sharp and Dr. Bruce Draper for supervising me to finish this project. They are the professors that sparked my interest in research and led me towards the field of science. I would not overcome many obstacles in the research progress without their help and encouragement. I would like to thank the vision lab members Dr. James Pustejovsky, Rahul Bangar, Gururaj Mulay, Pradyumna Kumar, David White, Matthew Dragan, Dhruva Patil, Nikhil Krishnaswamy, Joe Strout, Jason Yu, and many previous students, the project would not be established and continuously refined without their contributions. I would also like to thank my mother Dongxing Zheng for her support when I feel stressful. I would like to thank Elliott Forney, who created and supported this LaTeX template. Although I have never met him, his work was invaluable in the creation of this package and has helped many students get their thesis approved by the CSU graduate school. This work was funded by the DARPA Communicating with Computers program under ARO contract W911NF-15-1-0459.

DEDICATION

I would like to dedicate this thesis to my mother Dongxing Zheng.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction

User Interfaces take a variety of forms, one of them being widely researched in recent years is an ultimate interface of a virtual person. For example, some scientists created a 3D virtual environment and a virtual agent that interacted with objects to emulate a real-world scenario. Specifically, when the users rather than solely computers/algorithms have the eligibility to control the virtual agent, the latter is called an avatar.

This thesis presents an empathic avatar named Diana that recognizes human affect and responds with natural facial expressions to improve user experience in the interaction. In previous studies in Human-Computer Interaction, many empathic agents only have the ability to make a conversation with the user (they are called embodied conversational agents), e.g., communicate with users in terms of dialogues and react with behaviors during the conversation. But they can not recognize natural gestures from naive users and collaborate with the user to finish a task. In this thesis, we extracted socially non-behavioral cues from human pairs that working on building tasks in a blocks world and referenced their affective metric relationships as guidance to design our avatar.

In the Human-Avatar Interaction, Diana was engaged in a problem-solving exercise with a user, the task for this empathic avatar was to work with the person to build structures (a staircase, etc) out of virtual blocks in a blocks world. Communication between a user and Diana included both non-verbal communications like gestures, eye gaze, and facial expressions, and verbal communication, e.g., speech. Concretely, in our system, the movements of the user's facial action units were considered as spontaneous signals and were interpreted as emotions, they were treated as signals that Diana was designed to adapt to and further improve. Consequently, this is one of the few studies to date investigating the role of affect in task-focused Human-Avatar interaction.

Diana's behaviors were designed based upon the observed behaviors of human dyads that collaborated on a task to build wooden blocks with gestures and/or conversations. A dyad defined a pair of human subjects in our experiment that worked on the same task. For each dyad, the two human subjects were separated into two different rooms and connected via video communication. One of them who worked as a *signaler* was given a block structure pattern, and he/she needed to give gestural or verbal instructions following the pattern to guide the other person who acted as a *builder* to build blocks. Their interactions were recorded into individual video and all the footage composed a video dataset called the EGGNOG (Elicited Gigantic Library of Naturally Occurring Gestures) [2]. In EGGNOG, we extracted the most frequently occurred gestures and used them on training the gesture set Diana could recognize. We also analyzed the affect relationships between *signalers* and *builders* to model the affective states of Diana.

Our work added to Diana the ability to recognize and express human-like affect in simulating empathy. Affect has been long studied in the field of Human-Computer Interaction. One of the main research orientations is Affective Computing, an interdisciplinary field spanning computer science, psychology, and cognitive science that involving methods to recognize, interpret, process, and simulate human affect using computer systems/algorithms [4]. One of the motivations for the research is the ability to give machines emotional intelligence, including to simulate empathy [4]. While there are many studies discussing the advantages of using embodied agents in recent years, some researchers point out that embodied agents can also increase some users' anxiety and effect users' interaction experience [5]. This finding reveals the importance to design human-centered embodied agents that can improve user perception and experience. Concerning about the perceived load of humans in a case that the avatar and the user collaborate to finish a task, we equipped Diana with the ability to coordinate with the user's affective states, such as expressing empathy like humans companions when the user feels negative emotions.

This thesis aims at providing an empirical human subject study of Diana with the ability to simulate empathy. To compensate for the shortage of studies of empathic agents in the CS field [6], we used psychological theory background as the ground-truth guidance and conducted an interdisciplinary investigation. Diana's gestures were modeled on human naturally occurred gestures, and her facial expressions were synthesized based upon human affects and psychological concepts from Yacin's hierarchical model of empathy [6]. Our study added another step to the research of natural 3D user interfaces in providing a human-centered experience.

In our work, affect essentially means Diana's ability to recognize user facial expressions as emotions and generate empathic facial expressions. In Diana's affect module, the actual user emotion detection was implemented by an expression recognition toolkit called Affdex originated from the Affectiva Team [7]. Diana's facial expressions were synthesized by designated combinations of action units and controlled by the linear movements of facial morph targets. Different from the seven basic emotions published by Ekman, the synthesized facial expressions on Diana's face could convey affective states such as concentration, confusion, joy, and sympathy. The mapping mechanism from facial expressions to affective states referenced two previous works and the well-known Facial Action Coding System [1] and combined with our addition of some action units.

1.1 Diana System and Her Basic Functionalities

Considering a scenario of finishing an assembly task by controlling an avatar to execute multiple steps in a virtual world. Modalities can be involved in this Human-Avatar system including verbal communications such as speech, and non-verbal communications like eye gaze, postures, gestures, and affect, specifically facial expressions. The advantage of such a multimodal avatar is it can both see and listen to instructions from the user. Unlike Siri [8] or Alexa [9], the system provides users a 3D environment that emulates the real-world and an embodied agent interacts with surrounding objects, thus it is more efficient and provides a human-like perception.

Our Diana system is one of the state-of-the-art multimodal intelligent systems. The system was a joint creation of James Pustejovsky's lab at Brandeis University and the CwC Lab at the Colorado State University. The external equipment of this system included a laptop, a desktop monitor for projection, a Microsoft Kinect v2 sensor [11], an HP 4310 webcam, a Yeti USB microphone, a keyboard, and a mouse. Figure 1.1 showed the setup in our lab when one of the researchers was giving gestural instruction to train an earlier version of Diana. A rectangular interaction zone was

[1]Figure used with permission from the authors in [10].

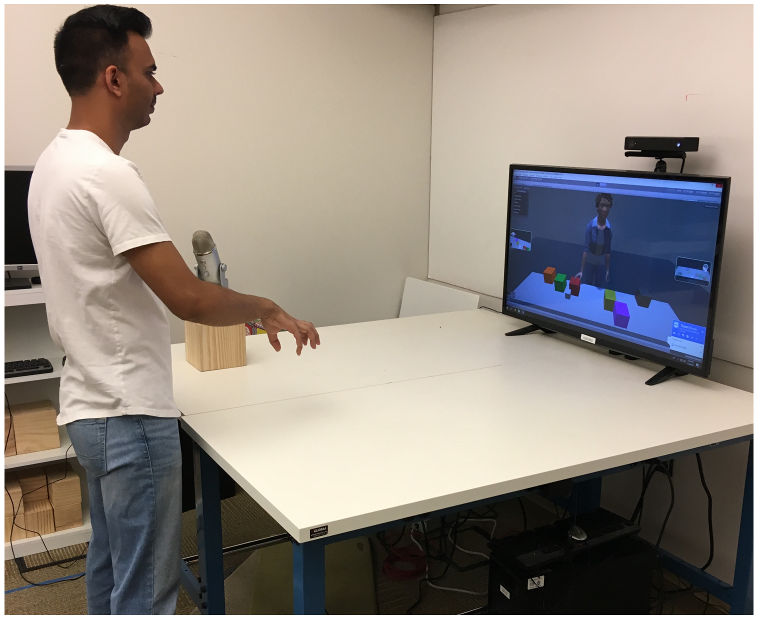

Figure 1.1: Lab setup used the Kinect v2 sensor for training[1].

bounded by blue tapes on the ground in front of the table, and it divided out the area (1.6∼2 meters, nearest to farthest, -0.8∼0.8 meters, left to right) where the sensor monitored user activity. Once the user stepped into the interaction zone and gave a wave, Diana would say "Hello, I'm ready to go." and awaited the user's next gestural or verbal instruction.

Figure 1.2 presented the internal framework and interface of Diana system. The whole system was developed in one Unity project. The virtual scene in the right part of the window was called the BlocksWorld, which depicted a scene that the user and the avatar collaborated with each other to build virtual blocks on the table. During interactions, Diana's arm motions, verbal responses, and facial expressions could all be seen by the user in BlocksWorld. To guide Diana's action, the gestures users were allowed to use are: waving (to attract Diana's attention at the beginning of a task), pointing (to select a certain block or a location), pushing (to move a block to the side of the table), servo (to move a block a little) and never mind (to undo Diana's last action). Verbal instructions Diana could recognize were the words referring to actions, prepositions, or colors, and semantics such as "put the red block on the green block", "put the yellow block to the right of the blue block". Users could also say "never mind" to undo her last action. Diana also reacted to another non-verbal instruction which was the user's head pose. The user's head pose was estimated

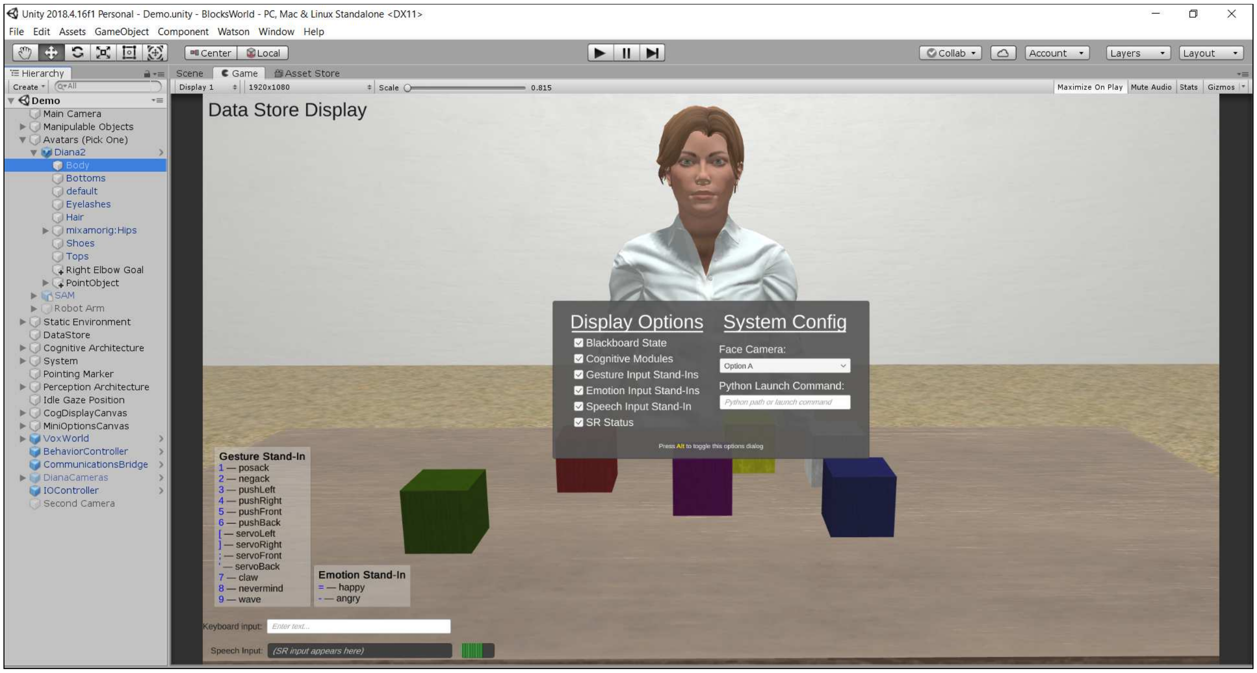

Figure 1.2: The framework and interface of Diana system.

by the Kinect sensor [11] that had a configured threshold of Euler angles. Once the user turned head left or right and exceeded the threshold, Diana would mirror the user's action and face the same direction where the user is looking at.

Including objects in the virtual environment such as the camera, the blocks, and the avatar, the system's functionalities were implemented in a hierarchical architecture that was shown on the left area of Figure 1.2. "Datastore" worked as the fusion that monitored and stored the key-value pairs that were actively updated by other modules. For example, a key "user:joint:Head" stored a Vector3 value that represented the location of the head point of the closest body frame in the "camera space". The "Cognitive Architecture" included modules that processed all the inputs (RGB-D images, verbal instructions, etc) from the Kinect sensor [11] and microphone. The architecture also had built-in mechanisms that controlled Diana's behaviors like blinking, arm motion, and generated speech responses. The "Perception Architecture" included modules and APIs for interfacing the webcam, capturing the user's skeleton data, arm motion, and hand poses into RGB-D frames that to be sent to the "Cognitive Architecture" for later processing.

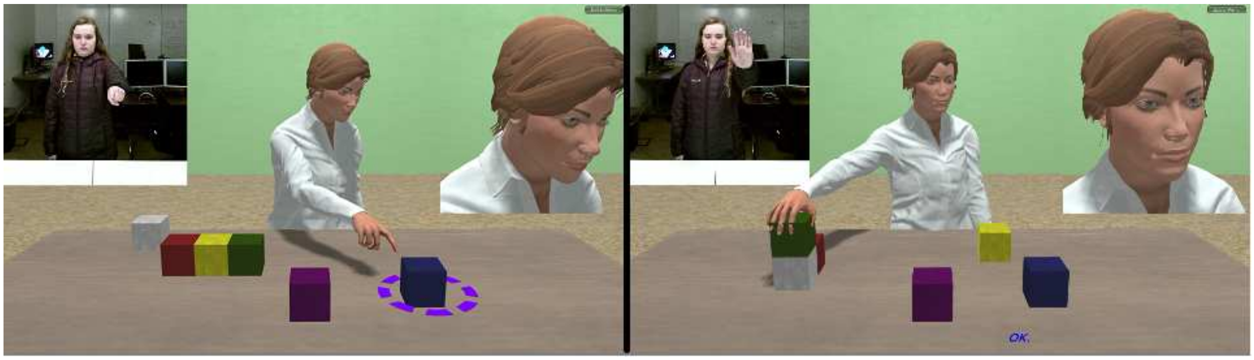

Figure 1.3: Pointing and never mind, the two most frequently used gestures during user interaction.

Speech recognition was implemented by using the Google Speech-to-Text Engine [12], this technology provided fast textual responses on speech recognition when dealing with multi-cultural accents. A C# class of Affdex SDK for affect recognition was attached under the "Perception Architecture". As complementary functionalities, The "CognitiveDisplayCanvas" was used during the developmental process for monitoring the values associated with keys in each module, it could display the key-value pairs in real-time on the left side of the window. Finally, the "MiniOptionsCanvas" was a panel that allowed the researchers to switch between different external cameras for face recognition, and enabled/disabled the display of the status of each module.

Figure 1.3 showed a human subject showing the two most frequently used gestures: pointing and never mind. Concretely, to select and then move one block, the user needed to point towards the screen. To help with the user to understand the exact location he/she was pointing at, a purple circle was displayed on the table surface as a pointing marker and it moved simultaneously when the user was moving his/her finger. Once the pointing marker overlapped with one block for several milliseconds, the marker's location was recorded by the system, and the block at that location was grasped by Diana. As shown by the right sub-figure, to show a never mind gesture, the user was required to keep palm vertically facing the screen with all fingers closed.

One characteristic we pioneered in our system was asynchrony. For example, during grasping, as soon as Diana heard verbal instructions like "No, the red one", or she had recognized a never mind gesture, she would immediately modify her actions without further instructions or gestures needed from the user. In the meantime, the marker on the table also kept tracking the user's pointing location until it located another stable spot the user was pointing at. Then the system passed the location into Diana's arm motion controller to invoke her to move the block to the final location. This asynchrony feature avoided the traditional listen-execute process in many interactive systems, thus reduced the time of completing an assembly task by quickly responded to the user's actions. The system efficiency and user engagement could be improved.

1.2 Modeling Diana's Behavior upon Human Behavior

Including the aforementioned gestures in the previous section, there were 32 different left/right-hand gestures Diana recognized, and they were all extracted from EGGNOG dataset [2]. The dataset included over 7 hours of RGB video, depth video, conversations, and 3D pose estimation data of 30 human dyads. It is a rich resource for evaluating naturally occurring gesture recognition systems. The hand recognition was implemented by the ResNet deep learning framework [13]. For estimating hand pose, we wrote Python clients to take the frame information (such as the depth data provided by the Kinect sensor [11]) as input and generated byte arrays then sent them to fusion. In our project, the HP 4310 webcam worked as a separate channel that connected with the Affdex SDK [14] to process RGB frames without the depth data.

1.3 Affect Recognition Using the Affdex SDK

Affectiva [7] is a human perception AI company that originated from the MIT Media Lab. It provided a multi-platform SDK named Affdex [14] for developers. Unfortunately, in 2019, the Affectiva Team no longer made their SDKs available to developers or academic research outside of the Imotions platform (a software platform combined with biosensor to aid behavioral human research). In previously released versions, Affdex could capture and process video streams from the camera or videos, and output spreadsheets including timestamp, perceptions of human age, gender, ethnicity, as well as 7 basic emotion metrics (joy, fear, disgust, sadness, anger, surprise, and contempt), 20 facial expression metrics and 4 appearance metrics (valence and engagement, etc). These metrics were trained and tested on over 6 million facial videos from more than 87 countries, representing real-world, spontaneous facial expressions made under challenging conditions. The key emotions could achieve accuracy in the high 90th percentile. An overview of the emotion AI can be found at their webpage: https://www.affectiva.com/emotion-ai-overview.

1.4 Facial Expressions Generation

Diana's dynamic facial expressions were synthesized based upon studies of action units from two previous works and combined with our additions. A particularly interesting insight was the work from researchers in the iVizLab at the Simon Fraser University [6]. Their work first provided a review of the empathy research from various fields, then proposed a hierarchy model of empathy that could be integrated into an interactive conversational agent. The model was composed of three layers: communication competence, emotion regulation, and cognitive mechanisms, from low to high. Communication competence meant emotion recognition, expression, and representation. Emotion regulation represented the self and relationship-related factors such as mood, personality, and affective link between the user and the agent. Cognitive mechanisms included perspective-taking which meant the agent thought from the user's perspective, and the Appraisal theory that the agent evaluated the environment and then gave appropriate responses. At last, the researchers summarized that existing models and implementations of empathic conversational agents lack the competence for the model presented in their paper, which indicated in the Human-Computer Interaction field the research on empathic agents still needed to be explored.

In our developmental process, we also tried to increase the user's recognition accuracy and their judgments of human-likeness of Diana's facial expressions by referencing the findings from Chen et al. at the University of Glasgow [15]. Inspired by their results, we selected facial regions around eyebrows, nose, cheek, and lip corners on our avatar's face to linearly manipulate, and utilized the action unit definitions in Facial Action Coding System [1] to synthesize them into different expressions of affective states.

1.5 Experiments and Results

Two experiments were conducted in our study. In the first, videos from EGGNOG [2] were processed by a C++ application provided by the Affectiva Team in their GitHub repository: https://github.com/Affectiva, and CSV spreadsheets that contained user information and affective metrics were generated. To inspect how the increase/decrease of human affects influence each other in each human dyad, we plotted the valence scores of the *signaler* and the *builder* on the same timeline, and then merged the plot with the videos that recorded their interaction. Merged results indicated that there was no correlation between the positive or negative nature of affect within a human dyad. Overall, a plot of all the interaction tasks showed the *signalers* had a larger range of valence than the *builders*, which meant their affect were more varied than their collaborators.

Considering the score samples were not normally distributed, we ran non-parametric statistical tests on the valence and emotion metrics. We found that when a user was assigned the role of *signaler*, the *signaler's* valence scores were significantly greater than the scores when the user acted as a *builder*, and the scores of joy in a *signaler* was also greater than the *builder*, indicating a happier *signaler* than a *builder*. This finding provided us inspiration to create a joyful expression of Diana when the user was happy, and reminded us to avoid letting Diana mimic the same negative facial expressions as the user to prevent exaggerating the bad mood.

In our second experiment, we aimed at investigating whether an empathic mode of Diana with dynamic facial expressions could improve user perception and experience, and whether such an avatar was preferred by users over the Diana with only mimicry ability or a flat face. Figure 1.4 showed the experimental scenario of the Mimicry mode of Diana in our user study. Diana was showing a joyful facial expression when she perceived the user was happy. To help users better observe the changes on Diana's face, a zoom-in camera view of Diana's head was simultaneously displayed at the top-right corner of the screen. For research purposes, a video stream that recorded user interaction was displayed at the top-left corner of the screen. The facial "BlendShapes" (which were actually morph targets on Diana's face that could be deformed to pre-defined shapes) were surrounded by red lines in Figure 1.5. Due to software limitations in the Unity platform and the

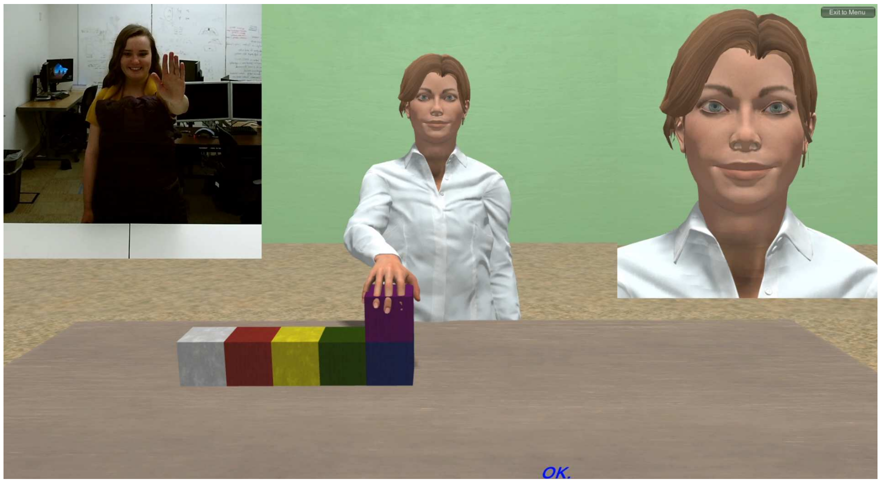

Figure 1.4: The mimicry Diana shows a joyful facial expression when she perceives the subject is happy.

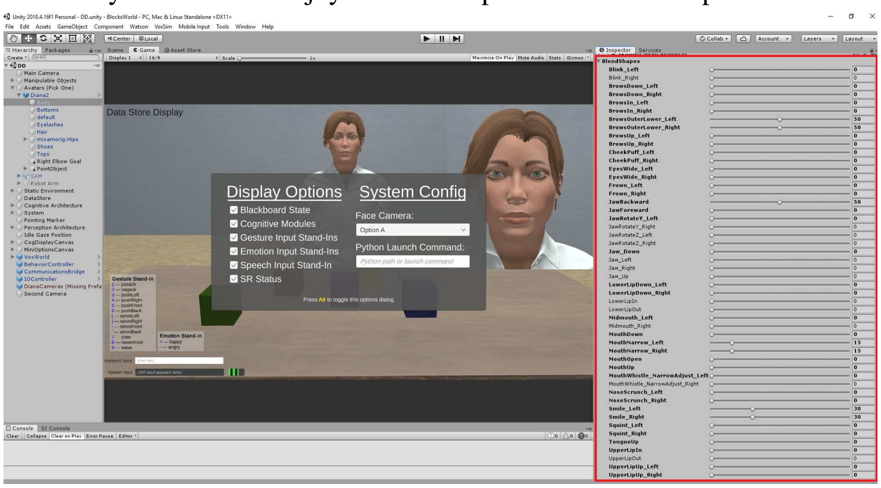

Figure 1.5: The action units we can control on the character.

Maya character, these facial morph targets merely controlled a subset of all the action units defined in the Facial Action Coding System [1].

During the experiment, users needed to collaborate with all three modes of Diana and finish tasks of building blocks. The three modes were called Emotionless (Diana with a flat face), Mimicry (Diana mimicked users' instant facial expressions), and Demo (Diana simulated empathy and expressed affective states through transitions of facial expressions), all other functionalities or appearances associated with these three modes of Diana were exactly the same. All the 21 human subjects had interacted with each mode three times and then gave votes in a 5-point Likert scale perception questionnaire and a NASA Task Load Index survey [16] right after the interactions.

Subjects length of completion of each task was also recorded automatically by the system logger in seconds.

Though we did not find significant variability in the votes or length of completion between the three modes of Diana, mode Demo received more positive votes (Agree and Strongly Agree) in 5 out of 7 questions in the perception questionnaire. Users also spent more time with the avatar in mode Demo than mode Mimicry or Emotionless. In subjective reports, users left comments for the Mimicry and Demo Diana saying they felt the facial expressions were natural and friendly, and chose these two modes as their favorite.

Chapters in this thesis are organized as follows: Chapter 2 describes previous works of embodied agent with affect that related to the work presented here and summarizes findings from the literature that using affect in the context of paired task-driven human interaction, it also presents the literature of affective embodied agents that inspired our design of the empathic Diana. Chapter 3 introduces the creation of EGGNOG [2] and a detailed view of Affdex [14] software we used to process our data, then reports a descriptive analysis and statistical tests ran on the affective metrics. Chapter 4 illustrates the experimental setup of our human subject study, with our results and analysis of subjective and objective data, following by the discussion of limitation of our study. We conclude in Chapter 5 with some suggested new avenues of research aim at developing better empathic agents.

Chapter 2

Literature Review

This chapter summarizes previous works in three sections including the studies about integrating human affect into embodied agents, modeling empathy by analyzing human psychological behaviors, and presenting the challenges researchers have encountered and some potential solutions.

2.1 Affect in Embodied Agents

Previous studies had found the value of including non-verbal communication channels in embodied agents [17], the audiovisual signals made it easier for users to perceive the internal state of an interactive system. For example, projecting uncertainty in a Question&Answering system. Embodied agents with emotional intelligence were considered more human-like, engaging, and trust-worthy [18–21]. Compared with emotionless agents, users rated higher subjective scores when they worked with affective agents [22]. They also spent more time interacting with these agents and indicated they were more willing to use the agent in the future interactions [23].

Affective Computing is one important method when designing the model of an affective agent. To describe a widely accepted prediction in Affective Computing, Pantic and Pentland [24] proposed the concept of human computing, indicating that anticipatory user interfaces should be human-centered, built for humans and based on human models. They concluded that human behaviors such as affective and social signaling were complex and difficult to understand, but these natural interactive signals also had much potential.

Regarding analyzing human affect between human interaction, many researchers have studied the recognition and perception of human emotions during two-way linguistic interactions. To detect emotions in the context of automated call center services, Devillers and Vasilescu [25–27] annotated the agent-client dialogues, and validated the presence of emotions via perceptual tests. Their first study included two types of emotion annotations. In the first type, two annotators independently listened to dialogues and labeled each sentence with one of the five emotions. In the second type, the annotators labeled the sentence without listening to the audio signal or the dialogue context. Then an emotion detection model was trained on the annotated corpus. Their second study carried out two perceptual tests with and without listening to the audio signal. Forty native French subjects participated in these tests. Researchers finally concluded that for accurate emotion detection, lexical, prosodic, voice quality, and contextual dialogic information needed to be combined.

Similarly, to detect the intensity of emotion felt by the speaker of a tweet, Mohammad and Bravo-Marquez [28] used a technique called Best-Worst Scaling [29] to create an emotion intensity dataset. The dataset provided reliable fine-grained intensity scores and was partitioned into training, development, and test sets for a competition. These researchers summarized the machine learning techniques and resources used by the participating teams that were particularly useful for determining the intensity of emotion. Twenty-two teams submitted the results of the shared task. It was found that affect lexicons, especially those with fine word-emotion association scores, were useful in determining emotion intensity.

There were also examples of work closer to that presented here in which two people collaborated to finish a task. To design affective interactive systems, Zara, Maffiolo, Martin, and Devillers [30] presented a protocol for the collection and annotation of multimodal emotional behaviors (speech and gestures) occurred during human interactions in a word-guessing game. Their experimental environment settings were similar to EGGNOG [2]. In their experiments, a human dyad was composed of a naive subject who guessed the word and a confederate subject who described the word. The confederate subject was asked by researchers to look at the word and hint on every card, and he/she needed to intentionally elicit a list of emotions from their naive partners, but the confederate subject could not mention five forbidden words given by the researcher. Naive subjects were 10 university students and confederates were 8 close relations of the experimenter or laboratory staff. The corpus between subjects was then analyzed from the viewpoints of third human judges. At last, the researchers illustrated the richness of the dataset with respect to expressions of emotions and other anthropomorphic characteristics.

Marsi and Rooden [17] summarized that previous studies about human-human dialogues found that non-verbal means such as speech prosody, facial expression, or gesture were used as cues to estimate the level of certainty. In a multimodal Question&Answering system, subjects judged the linguistic signaling of uncertainty worse than their visual cues counterparts. Researchers also argued that expressing uncertainty was related to trust, and believable embodied conversational agents were an important factor in building trust relations between humans and computers. To justify that in a multimodal question-answering system certainty/uncertainty can be reliably expressed by means of animated facial expressions. Fifty subjects watched a sequence of web pages of an agent's head and finished 5 point Likert scale questions. Results suggested that humans could correctly recognize certainty through the animated head's facial expressions. Either eyebrow or head movements were sufficient to express certainty. However, only head movements and combined movements significantly expressed uncertainty. In contrast, eyebrow movements were perceived as signaling certainty.

With the development of natural user interfaces in Human-Computer Interaction, virtual agents with the ability of affect started to become a hot spot in research. Scientists like Ku et al. [31] investigated how a human affectively perceived an avatar's facial expressions. In Ku's work, a male and a female virtual avatar with 5 levels of intensity of emotions were generated using the morphing technique and were displayed to 16 graduate students. Researchers found that as the facial expressions displayed on the avatar's face became more intense, subjects were evoked to have higher values of affective valence and arousal. Their finding exemplified that an avatar with a facial expression of a certain level of emotion could influence an experimental subject. In comparison of the responses to two genders of avatars, the male and female avatar evoked different incremental/decremental slopes in valence values when happy/angry intensified, but there was no significant difference between their arousal values. Their work also provided evidence that the intensity of emotions of an avatar could be controlled by linear morphing of facial expressions.

At last, researchers also discussed the limitations that though subjects could recognize the avatar's facial expressions well, they were not emotionally affected to the same extent because they might think the avatar was not real.

Some researchers studied the influence of letting agents mimic the user's emotions and facial expressions. In the work of Shen et al. [32], the influence of sentiment apprehension by robots (i.e., robot's ability to reason about the user's attitudes such as judgment or liking) was analyzed in two conditions. In the first, the robot only mimicked user affective states, while in the second it also provided sentimental feedback of users' facial expressions. Thirty-two participants were randomly assigned to the control group or the experimental group to play a "Mimic-Me" game for as many rounds as they liked. The researchers found users spent more time interacting with the robot that had the ability to understand the sentiment and gave higher ratings on this robot and concluded this robot rendered the Human-Robot Interaction experience more engaging.

As part of the work on a multimodal animated avatar, a study by Pablos et al. [22] presented a computational model that achieved high accuracy of facial emotion recognition with streaming videos as the input. These researchers built active shape models and Gabor filters in the action units recognition module and then fed results into a hybrid neural network. Subsequently, they integrated the model into an empathic avatar that randomly nodded when the participant was speaking and mimicked the participant's facial expressions. Twenty participants took part in the preliminary experiment and were randomly distributed into the condition of facing the empathic avatar, or the condition of facing an avatar responding to interaction with only a pre-programmed behavior. Results indicated the empathic avatar with the facial expression model received an increase in the positivity of users' ratings.

Similarly, Aneja, McDuff, and Shah [33] built a high-fidelity embodied avatar that could map human action unit movements to lip-syncing, head gesture, and facial expression capabilities. The avatar was controlled by its bone positions, phonemes, and action units. To avoid conflicts with lip movements it only mimicked the user's facial part above the lips region. Two pipelines were introduced in their paper. In the first, researchers tracked the human face in videos with a detector and trained a 3D Convolutional Neural Network to recognize 12 action units then synthesized similar facial expressions on the avatar. The second method implemented expression synthesis on the avatar's face via bone position controls. Though there was no user perceptual test performed on this avatar, the researchers released their code and model to the public to encourage research on creating conversational agents using these APIs.

Another study on affective tutors by Mudrick et al. [34] aimed at investigating how a tutor's facial expressions could influence learners' performance and emotions. Researchers used the Emotions as Social Information model and Dynamics of Affective States Model to explain the influence of the human tutor agent's facial expressions on the emotions and learning outcomes. Forty-four undergraduate students were asked to read diagrams and text related to Biology concepts in a MetaTutor Learning Environment with a human tutor agent's face displayed at one side of the window. The human tutor agent remained a neutral face when subjects were finishing reading the text and diagram, and expressed her facial expressions regarding the relevancy of content after subjects had made content relevancy judgments. It was described that human learners' performance was significantly better when a human tutor agent expressed facial expressions congruent with the learning content, and learners felt more confusion when the agent provided in-congruent facial expressions. The results had important implications for contextual congruency of virtual tutor emotion expressions in other contexts, such as mimicking learners' facial expressions to let them aware of their emotions, or representing the tutor's appraisal of learner's actions. Their findings supported our idea that controlling the agent's facial expressions can influence user perceptions.

2.2 Empathy in Embodied Agents

As mentioned in Section 1.1, our avatar modeled from human behaviors aimed at simulating a more complex affect empathy to improve the user's perception in tasks. Though affect has been intensively studied in Affective Computing, it is still challenging to create an empathic embodied agent. Empathy has been defined in the scientific literature as the capacity to relate another's emotional state and assigned to a broad spectrum of cognitive and behavioral abilities [6]. Agents merely mimicking human facial expressions lies in the fundamental level of a hierarchical empathy model and only represents a low-level empathic behavior towards users [6], thus this behavior is not sufficient to fulfill the requirements in today's affective Human-Agent Interactions.

Some researchers created agents that could learn and analyze the user's context-dependent behavioral patterns from multi-sensory data and adapted the interaction accordingly. In their study, the agent expressed her empathy through her appraisal of the environment. Chen et al. [23] designed a learning software with an upper-body virtual agent on the side of the window to promote students' engagement and enhance learning. Fifty-two college students were required to read e-textbooks and finished exercises, during reading they needed to select their current emotions from 4 pairs of emotional states. The empathic agent expressed her empathy through animated facial expressions, gestures, and voices. For example, if the user chose one of the negative emotions, the agent expressed emotions such as sadness or worried and held hands in a fist to encourage the user to continue. After learning, students were asked to fill-in subjective questionnaires to rate their feelings of the agent. Results showed that compared to the neutral agent, the empathic agent effectively reduced the student's boredom, and participants were willing to spend more time with her.

Additionally, to use affect within a decision-making process to improve the performance and attraction of a non-expensive robotic agent, Esteban and Insua [35] proposed an affective model for autonomous robots that calculated through mathematical models to infer user actions and environment evolution. The agent selected and expressed from four basic emotions and was triggered behaviors reacting to the expected or immediate human moods. Sixty-two participants were randomly assigned to play the rock-paper-scissors game with the affective or the emotionless robot for as many rounds as they liked. Then they were asked to fill in a Likert scale questionnaire. Results indicated that the affective robot was more appealing than the emotionless one and led to longer interactions.

During the process of synthesizing agents' facial expressions with action units, Chen et al. [15] at the University of Glasgow found the standardized facial expressions previously thought to be

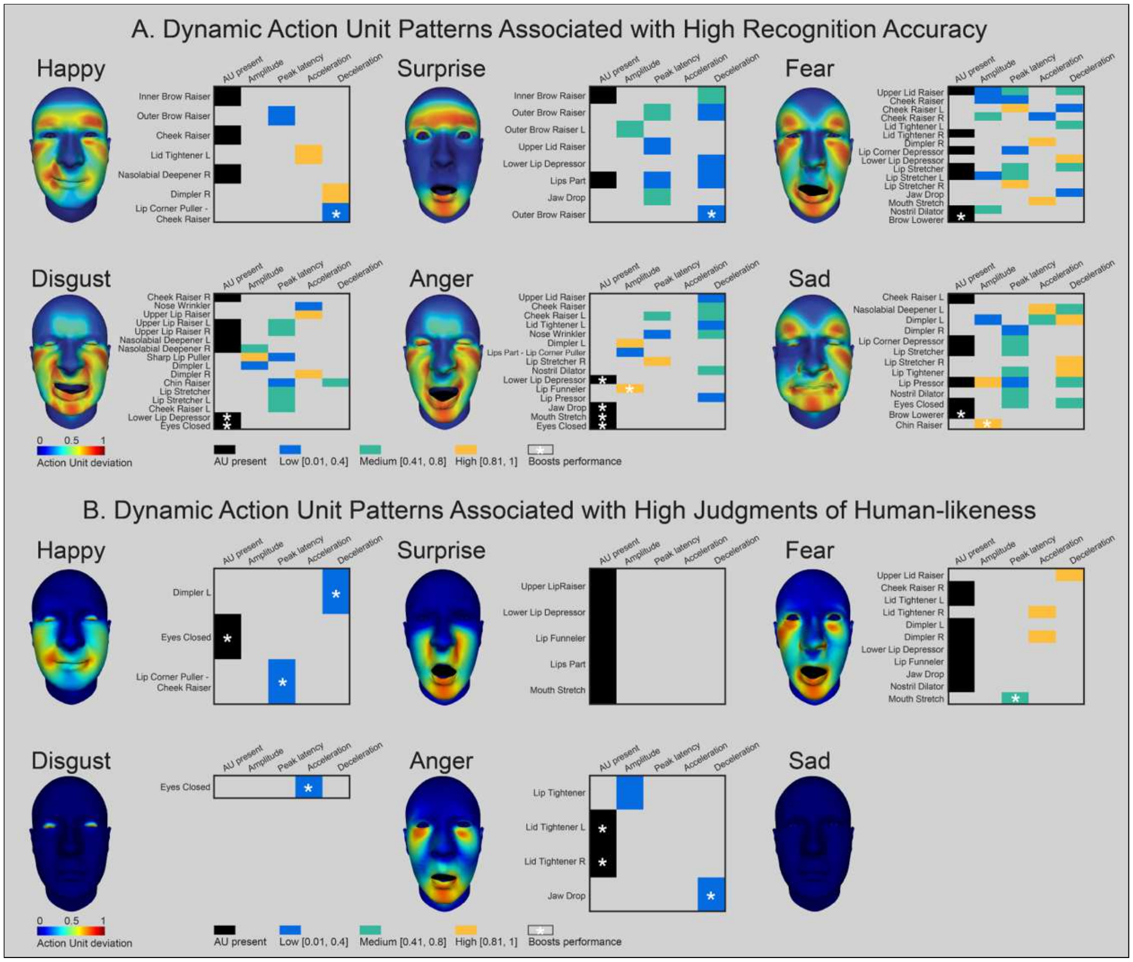

Figure 2.1: The culturally-sensitive dynamic action units that are associated with high recognition accuracy in panel A and high judgments of human-likeness in panel B (Figure used with the permission from the authors in [15].)

globally recognized were actually less accurately recognized in Asian cultures than in Western cultures. To develop culturally sensitive facial expressions, they generated random action unit combinations on the agent's face, then detailed the specific dynamic action units that were associated with high recognition accuracy or judgments of human-likeness, as adopted in Figure 2.1. In each panel, six basic emotions were investigated, the face maps showed the action units that were associated with high performance, the color-coded matrices also indicated any specific (unit interval) temporal parameter values associated with three levels of performance (low, medium, and high). Those action units that further boosted performance were indicated with white asterisks. Ten Chinese students were recruited in both two experiments. The first experiment asked about the

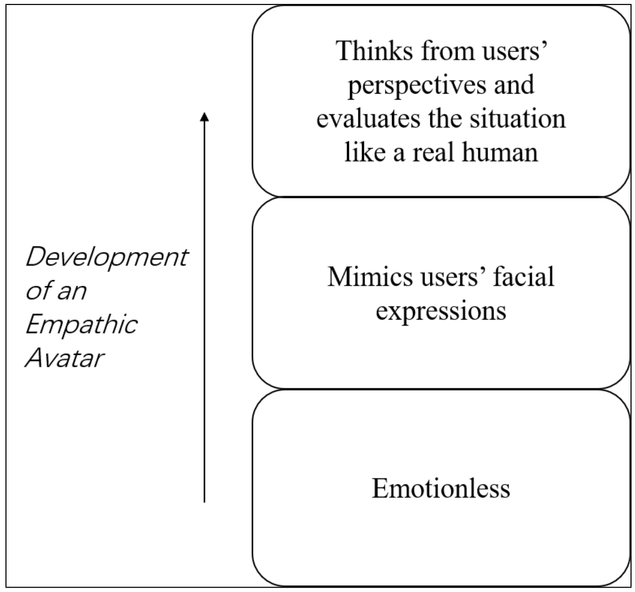

Figure 2.2: Three modes of Diana following a hierarchical model of empathy.

classification of emotion based on the facial action unit combinations on the agent's face, while the second experiment studied the judgment of human-likeness of the expression on the agent's face. Results showed that the modified facial expressions that take into account culturally-distinct responses were viewed more favorably in terms of accuracy and human-likeness than the standardized facial expressions.

We considered the findings mentioned in the paragraphs above and found that mimicking users' facial expressions was not sufficient to design a compelling human-centered avatar. We adopted the hierarchical model from Yalcin [6] and created three modes of Diana to give a comprehensive comparison of their ability of improving user experience, as shown in Figure 2.2. The three modes included a Emotionless avatar that was commonly studied in commercial activities and research, the Mimicry avatar with fundamental ability of affect that was studied by Human-Computer Interaction scientists, and the Demo avatar that thinks from the user's perspective with the ability of simulating high-level empathy.

I wish to acknowledge that the works at Simon Fraser University and the University of Glasgow were helpful and inspiring. The experiments presented here go beyond their work in so much are our work demonstrates affect in the context of multi-modal communication and task completion, but their prior work in general and in particular on affect generation guided our own work.

2.3 Challenges and Solutions when Adding Affect to Agents

There are also significant challenges associated with successively developing a working empathic agent. Cohn [36] clarified that in human-centered computing it was a mistake to think the goal was emotion recognition. Emotions were not directly observable but were inferred from expressive behavior, self-report, physiological indicators, and context. To make computers perceive, understand, and respond appropriately to human emotions without deliberate human input, it was argued that we forgot about the notion of "emotion recognition" but adopt an iterative approach found in human-human interaction. In his work, he included approaches to measurement, timing or dynamics, individual differences, dyadic interaction, and inference, and suggested that we consider the complexity of emotion when designing perceptual interfaces.

One of the challenges is the increase of user perceived load during interaction. In the study by Chen et al. [23], while most of the subjects expressed their willingness to continue interaction, some subjects also reported the empathic tutor had elicited more frustration and worry during the learning process. Another study by Haring et al. [37] illustrated that adding a robot's cheating behavior into a rock, paper, scissors game with humans had elicited more aggressive emotions in humans regarding the robot. Additionally, the interaction experiences with the robot were rated by participants as more discomforting compared to the experience with the human player.

Major challenges also include minimizing subjective perceptual differences upon the agent's synthesized facial expressions, especially the Uncanny Valley effect discovered in cognitive sciences. i.e., even subtle flaws in appearance and movement can be more apparent and eerie in very human-like but not identical robots, from the findings by MacDorman and Ishiguro [38]. To generate humanoids expressions, many researchers have provided diverse solutions. Belhaj, Kebair, and Said [39] had proposed an agent model that included emotions and coping mechanisms. The model emphasized the influence of emotions in the agent decision-making and action-selection processes and generated human-like behaviors for the agent. Scherer et al. [40] conducted a study to investigate some action units or their combinations that were most likely to be recognized as certain emotions under human appraisals. As defined in Wikipedia, Appraisal theory is a term in psychology describing that emotions are extracted from our evaluations (or estimates) of events that cause specific reactions in different people [3]. In Scherer's work, the results from three experiments involving 57 French-speaking students confirmed that participants could infer targeted appraisals and emotions from synthesized facial actions based on appraisal predictions. They also provided evidence that the human's ability to correctly interpret the synthesized stimuli was highly correlated with their emotion recognition ability as part of emotional competence.

The work by Rodriguez and Ramos [41] also illustrated the major challenges in building computational models of emotions for autonomous agents and presented a novel approach. The challenges included integration of cognition and emotions in agent architectures, unification of the various aspects of emotions, scalable architectures for computational models of emotions, and exploitation of biological evidence. Their approach was composed of a 3-layer integrative framework and a general methodology to guide the development of biologically inspired Computational Models of Emotions within 3 phases. They suggested that researchers taking advantage of theories and models from fields that study the brain information processing that underlies emotions, such as neuroscience, neuropsychology, and neurophysiology, and simulating the agent's emotional mechanisms based on these conceptions.

Chapter 3

Affect in the Context of Human-Human Interaction

This chapter first introduces the creation of our EGGNOG dataset, then presents a descriptive analysis and results from non-parametric statistical tests of human affective relationships in the videos from EGGNOG [2].

3.1 Dataset

Our EGGNOG dataset includes 7 hours of footage of naive user nature interactions, the videos were collected in an elicitation study conducted by Wang et al. in 2017 [2] at the CwC Lab and were made publicly available at: https://www.cs.colostate.edu/~vision/eggnog. We created this dataset aimed at providing a large-scale video resource to stimulate research in natural and gestural human-computer interfaces. The characteristics of EGGNOG include: 1) Naturally occurring gestures during human-human interaction. The participants were given no instructions about gesturing, thus their gestures were different from the pre-defined gestures in common user interfaces. 2) Continuous, Labeled Data. Each video contains the entire gestures and communications between two people. Gestures were semi-automatically segmented as motions (e.g., "RA: move, up; RH: into two, front;"), and the intents of the participant were also labeled (e.g., "two new

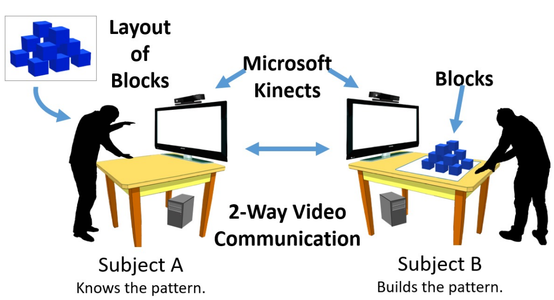

Figure 3.1: A human dyad is composed of a *signaler* and a *builder*[2].

Figure 3.2: A frame of the merged video of one dyad in EGGNOG dataset.

[2]Figure used with the permission from the authors in [42].

blocks"). 3) High-Quality Multi-Modal Data. This dataset was captured using a Microsoft Kinect v2 sensor [11] that outputting RGB videos, depth, and skeleton joint position data.

As shown in Figure 3.1 and Figure 3.2, two human subjects in separate rooms interacted through a two-way video communication to finish a task of building blocks, and their behaviors were recorded separately into two videos. Subject A who was assigned a pattern of block structure was called the *signaler*, while Subject B who was given several wooden blocks beside table was called the *builder*. The pattern with a layout of blocks usually required the user to stack blocks or move blocks, a typical one was a 6-block staircase with 3 blocks at the bottom layer, 2 blocks in the middle layer, and one block at the top. The goal of the *signaler* was to guide the *builder* to replicate the arrangement of blocks of the pattern using gestures and/or speech. In all the 30 dyads, 10 of them could see and hear each other, while another 10 dyads could only use gestures to communicate and the remaining 10 dyads could only converse. Patterns were randomly assigned to each dyad without replications. Each dyad was asked to finish 10 random building tasks first, then the *signaler* and *builder* exchanged their roles and finished another 10 random tasks. At last, crashed videos and incomplete tasks were marked in the dataset and removed from analysis.

This dataset informs the construction of our Diana system, a system that recognizes 32 human natural gestures that were extracted from EGGNOG [2] to aid the user. The dataset also provided us information on the human affective relationships in two-way human interactions, and thus guided us to design the empathic behaviors of Diana to let her act like the *builder* while the user acted as the *signaler*.

3.2 Recording Human Affect in EGGNOG

The design of an empathic avatar requires us to observe how humans affectively communicate with each other under a real-world scenario. To label human affects that appeared during human interactions, videos in EGGNOG [2] were processed by a configured version of the C++ application originally provided in the Affectiva GitHub repository: https://github.com/Affectiva.

TimeStamp	faceId	interocularDistance	glasses	age	ethnicity	gender	dominantEmoji	pitch	yaw	roll	joy	fear	disgust	sadness	anger	surprise	contempt	valence	engagement	smile	innerBrowRaise
0	nan	nan	no	unknown	unknown	unknown	unknown	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
0.0333	nan	nan	no	unknown	unknown	unknown	unknown	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
0.1	0	67.5346	no	unknown	unknown	unknown	Unknown	5.2394	-22.2026	34.5373	0.0018	0.0045	0.4105	0.0237	0.0023	0.1929	0.2134	0	0.1012	0.0258	0.0044
0.1333	0	73.7647	no	unknown	unknown	unknown	Unknown	-2.6378	-26.8587	-47.7827	0	0	0	0	0	0	0	0	0	0	0
0.2	0	70.5805	no	unknown	unknown	unknown	Unknown	-13.5201	-26.5721	4.7918	0	0	0	0	0	0	0	0	0	0	0
0.2667	0	70.7239	no	unknown	unknown	unknown	Unknown	-14.9689	-24.6357	5.5649	0.0018	0.0002	0.4385	0.0226	0.0022	1.3855	0.1938	0	0.0913	0.0011	0.0038
0.3667	0	70.3063	no	unknown	unknown	unknown	Unknown	-14.8599	-25.2612	7.9105	0	0	0	0	0	0	0	0	0	0	0
0.4333	0	70.9562	no	unknown	unknown	unknown	Unknown	-20.8327	-24.0446	5.2884	0.0019	0.0002	0.4862	0.0138	0.0028	1.51	0.1966	0	0.451	0.0025	0.0039
0.5	0	68.6048	no	unknown	unknown	unknown	Unknown	-18.5471	-30.8501	7.5308	0	0	0	0	0	0	0	0	0	0	0
0.5333	0	68.6893	no	unknown	unknown	unknown	Unknown	-20.6119	-29.5795	6.3808	0	0	0	0	0	0	0	0	0	0	0
0.6	0	69.4151	no	unknown	unknown	unknown	Unknown	-20.3694	-26.8762	10.0742	0	0	0	0	0	0	0	0	0	0	0
0.6667	0	69.4121	no	unknown	unknown	unknown	Unknown	-18.2766	-23.881	7.3049	0.0026	0.0029	0.6776	0.0116	0.0054	0.4512	0.2098	0	0.5545	0.0006	0.0056
0.7333	0	69.7874	no	unknown	unknown	unknown	Unknown	-18.551	-21.1393	5.3234	0.0025	0.004	0.6295	0.0127	0.0045	0.3416	0.2277	0	0.3987	0.0006	0.0057
0.8333	0	66.1087	no	unknown	unknown	unknown	Unknown	-19.7409	-20.0724	2.8569	0.0022	0.0043	0.5484	0.0163	0.0034	0.2721	0.2884	0	0.2295	0.0006	0.0058
0.8667	0	70.3079	no	unknown	unknown	unknown	Unknown	-17.6534	-16.2397	1.8093	0.0021	0.0046	0.5294	0.0177	0.0033	0.246	0.3007	0	0.2016	0.0006	0.0058
0.9333	0	73.6795	no	unknown	unknown	unknown	Unknown	-16.1497	-17.8829	3.3353	0.0019	0.0046	0.4607	0.0224	0.0025	0.2051	0.3158	0	0.1151	0.0008	0.0052
1	0	74.2691	no	unknown	unknown	unknown	Unknown	-18.2163	-9.5289	2.362	0.0018	0.0047	0.4554	0.0231	0.0025	0.1985	0.2736	0	0.1115	0.001	0.0048
1.1	0	74.796	no	unknown	unknown	unknown	Unknown	-18.9859	-4.5546	6.1182	0.0017	0.0047	0.4448	0.0245	0.0024	0.1896	0.2299	0	0.1031	0.0011	0.003
1.1667	0	74.3563	no	unknown	unknown	unknown	Unknown	-10.903	-1.2783	12.5523	0.0015	0.005	0.4356	0.0288	0.0027	0.1659	0.2494	0	0.1076	0.0009	0.0025

Figure 3.3: A partial sample output file of Affectiva.

When processing each video, all the affective metric scores of the subject were encoded and output in a time series in a frame-by-frame CSV file as shown in Figure 3.3. Columns from A to H recorded attributes including TimeStamp, faceId (the count of human faces appeared in this video, started from 0), and basic subject information: interocularDistance (the distance between two pupils), glasses (whether the subject was wearing glasses or not), age, ethnicity, gender, and dominantEmoji (the instant emotion that overwhelmed other emotions at that timestamp). The next three columns recorded the pitch, yaw, and roll of the subject's head pose in Euler angles. Followed by seven basic emotion metrics (joy, fear, disgust, sadness, anger, surprise, and contempt) in the scale from 0 to 100, and 4 appearance metrics such as valence and engagement. The mapping from facial expressions to emotions followed the Facial Action Coding System published and updated by Ekman et al. [1]. The metrics could be thought of as detectors: as the facial expression in a specific group occurred and intensified, the score increased from 0 (no expression) to 100 (expression fully present). In addition, valence scores from 0 to 100 indicated a neutral to a positive experience, while scores from -100 to 0 indicated a negative to neutral experience. Engagement measured the intensity of a group of the subject's facial expressions. At last, the movement of the key landmarks on the subject's face was encoded as facial expressions (e.g., smile) and also ranged from 0 to 100.

There were several anomalous aspects to the way the data was recorded. For example, in Figure 3.3, the first two rows was populated with "nan" and "unknown" values. This happened because when the subject's head turned beyond the maximal Euler angle the program could detect, "unknown" values were written to the user's profile and "nan" values were placed under all affec-

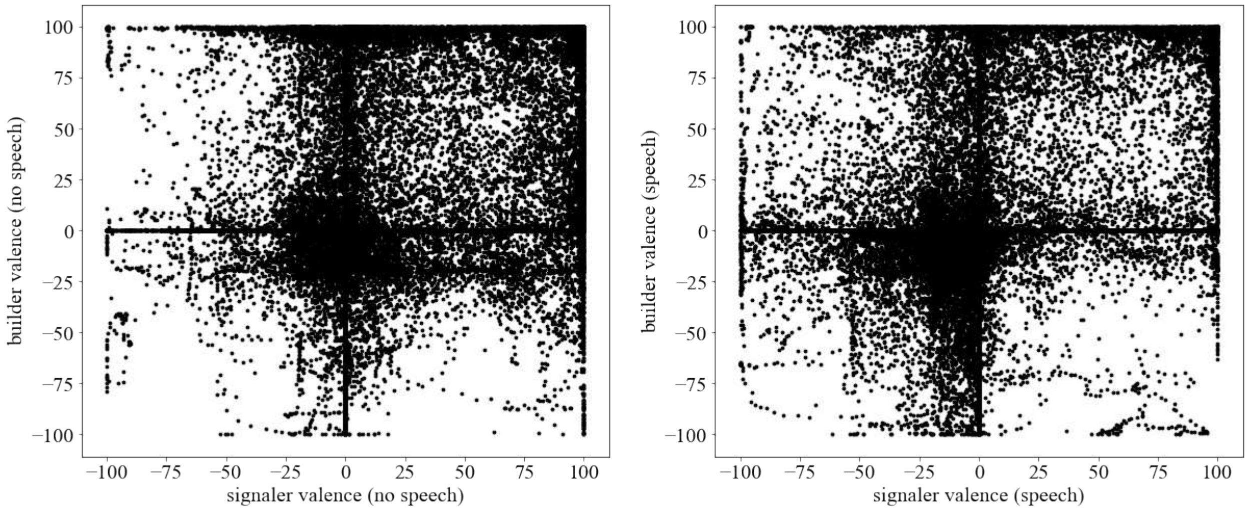

Figure 3.4: The scatter plot of *signalers'* and *builders'* valence w/o speech in video recordings.

tive metrics. In addition, when the software was unable to recognize any affective metrics from the subject's face, zero values were written to all emotions and facial expressions.

3.3 Data Analysis

3.3.1 Descriptive Analysis

As mentioned in Chapter 3.1, every video in the EGGNOG dataset [2] recorded one subject's behaviors and affects in one task. We cropped video frames into smaller rectangular with the subject standing in the middle to remove unimportant background information. Then we fed all the videos of *signalers* and *builders* into the application and obtained one CSV file per video.

To investigate whether the recognition accuracy of Affdex [14] on human affective states was influenced due to facial movements around lips, we separated results of emotion recognition in the videos with and without speech separately. The scatter plots in Figure 3.4 showed valence scores of all the *signalers* and *builders* in all tasks. The data points mostly located around zero and formed two horizontal and vertical crossing lines, indicating there was no linear or monotonic correlation between *signalers'* valence and *builders'* valence. When we observed the valence scores distribution with/without speech, the was not much difference, meaning that the recognition accuracy of affective valence was not influenced the user's lip movements.

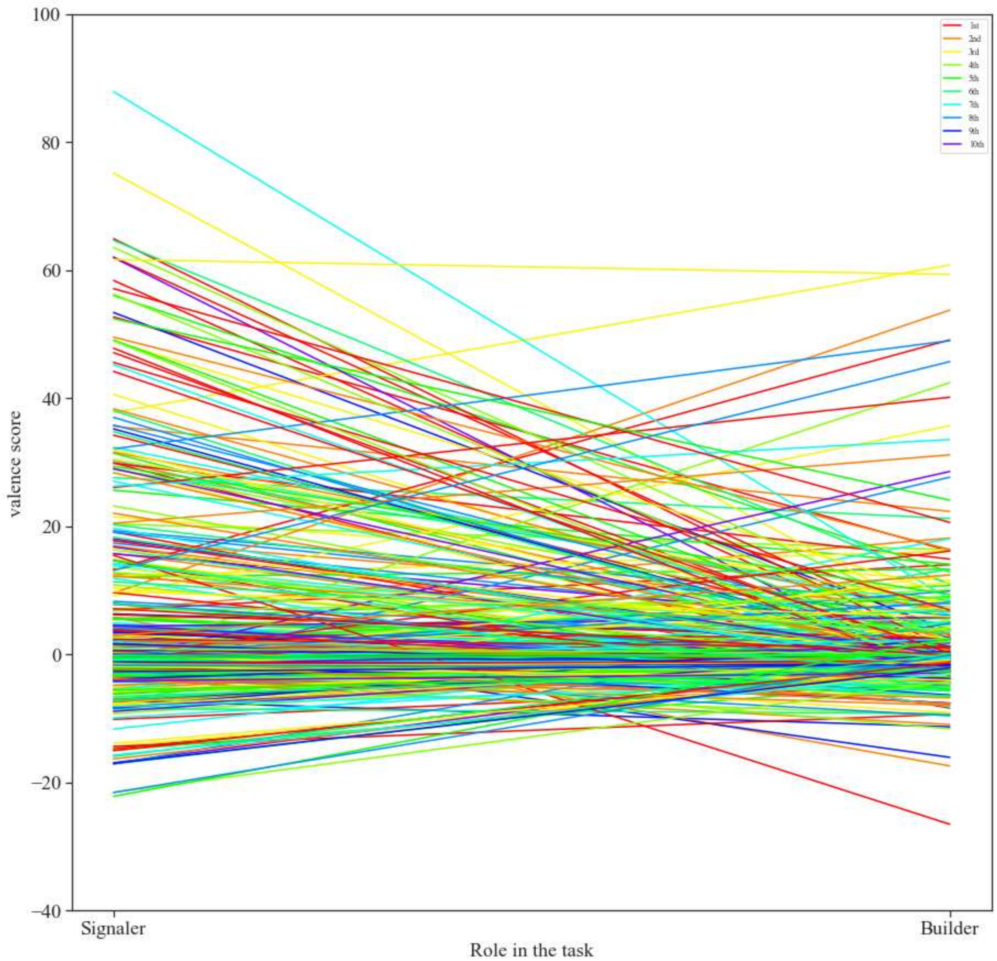

Figure 3.5: The overall scores of affective valence of the *signalers* and *builders* in all tasks.

We first inspected the overall distribution of affective valence between different roles of users. Each subject's valid valence scores in one spreadsheet were averaged (including zeros) and resulted in one score representing the participant's affective state in the task. In Figure 3.5, each line represented a connection between affective valence scores in a task of a human dyad, and the color denoted the order of the task. For example, the red line indicated the first task collaborated by a specific *signaler* and a *builder*, its left endpoint corresponded to the valence score of the *signaler* and right endpoint corresponded to that of the *builder*. Overall, the *signalers* had a broader range of valence scores than the *builders*, which meant they might experience more positive or negative emotions. Though there were many lines with nearly equal valence scores, those lines with negative slopes also had larger slopes in absolute values than those lines with positive slopes,

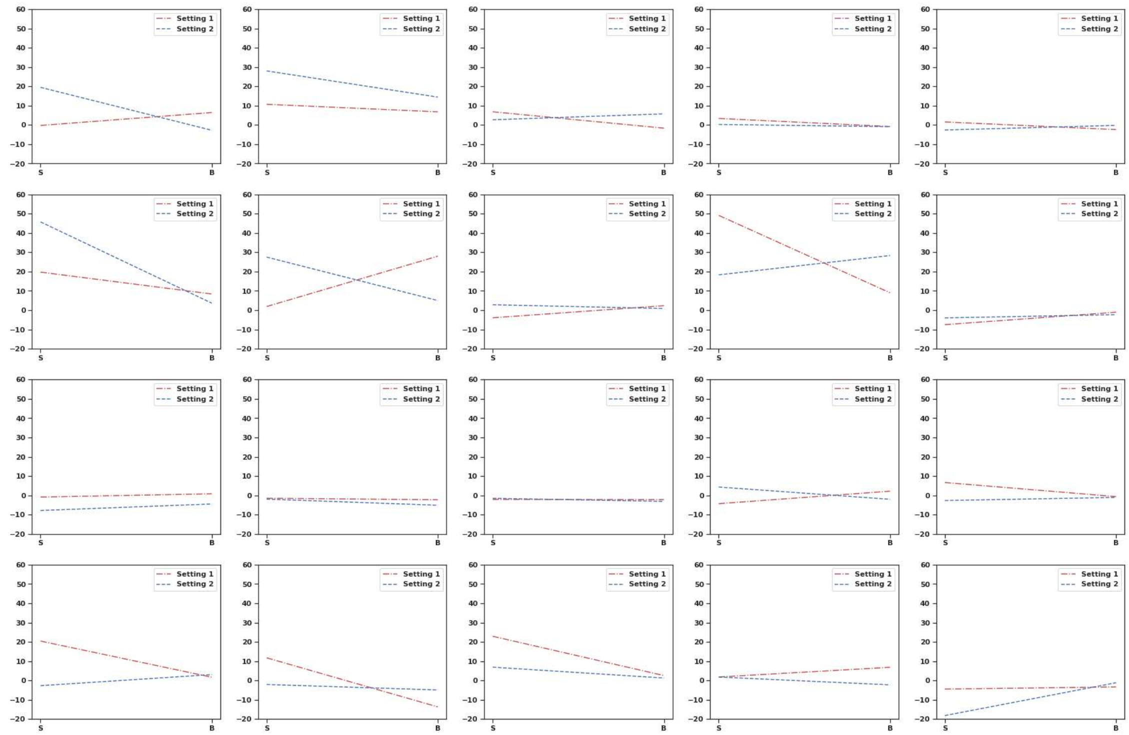

Figure 3.6: The scores of affective valence of the *signalers* and *builders* in their first task.

which indicated when the *signaler* and the *builder* were experiencing different valence of emotions, the *signalers* might feel more positive emotions than the *builders*.

We then analyzed the distribution of average valence in each dyad and investigated whether personality traits were related to the subject's role in a task and thus influenced valence. To reduce the confounding variables introduced by the learning process when users got familiar with each other in repeated collaborations, we only selected and analyzed videos from the first task carried out by each dyad. Figure 3.6 showed the distribution of valence scores in 20 dyads that with gestural information, for example, the red line in the first sub-figure represented the setting of subject A as the *signaler* and subject B as the *builder*, and the blue line represented the setting that they exchanged roles and finished another task. When the *signaler* and *builder* had noticeably different valence scores, in 17 cases the *signaler's* valence was higher than the *builders*' valence while in 15 cases it was lower. We admit this might happen by chance, but considering when the differences between valence scores were noticeable, in most subplots, it was the *signaler* that had a much higher valence than the *builder* rather not the opposite situation. A descriptive analysis of

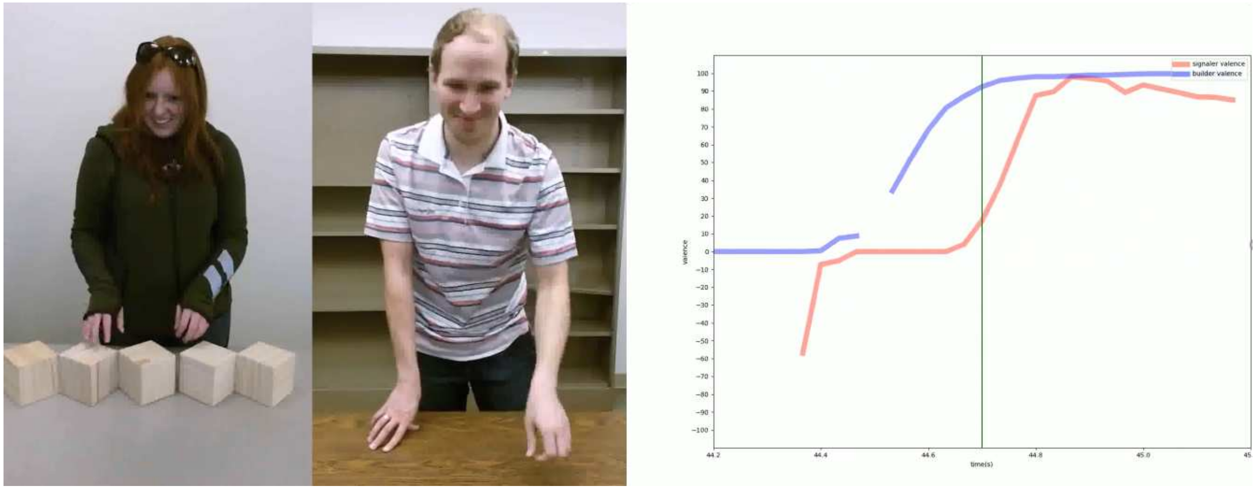

Figure 3.7: A frame from a merged video of a *builder* and a *signaler* and their corresponding plots of valence.

the valence scores of *signalers'* (*M* = 10.88, *SD* = 37.56) and *builders'* (*M* = 2.42, *SD* = 38.67) indicated the mean of *signalers* were larger than *builders*. We also found that the valence scores were associated with the role more than associated with the individual person, otherwise, in this figure we should see more paired crossing lines with similar slopes in absolute values, like the subplot in the second column of the second row showed.

We then inspected videos frame-by-frame combining with the plots. Figure 3.7 showed a snapshot at one certain timestamp of a merged video of a specific *builder* and a *signaler* finishing a task in the BlocksWorld and their corresponding plots of valence. The red line plotted the *signaler's* valence and the blue line represented the *builder's* valence, with a vertical green line at the center of the plot located the exact timestamp in seconds. Both the *signaler* and the *builder* were smiling during this time interval and their joy was indicated by an increase in reaching high valence scores (nearly 100). The breakpoints on the blue line were the missing values recorded when the female subject turned her face out of the camera frontal view. In this video, most of the time the *signaler* was smiling and trying to guide the builder. Although at certain points the *builder* showed confusion at understanding the *signaler's* behavior and her emotion was reflected by negative values, we also observed that she sometimes was influenced by the *signaler's* positive affective state and responded with a positive emotional response.

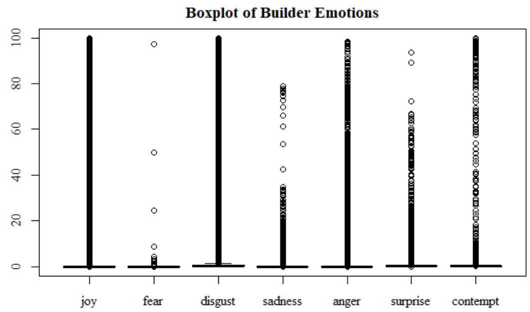

Figure 3.8: Box plot of *builder* emotion metrics.

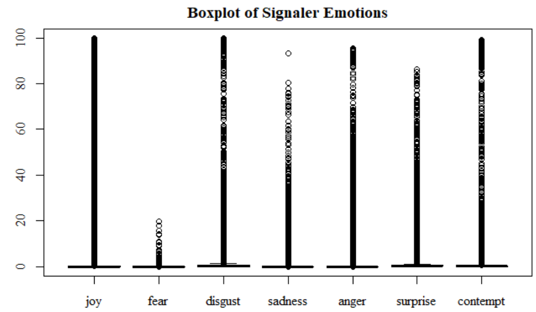

Figure 3.9: Box plot of *signaler* emotion metrics.

Table 3.1: A summary table of the *signalers'* and *builders'* emotions.

	signaler							builder						
	joy	**fear**	**disgust**	**sadness**	**anger**	**surprise**	**contempt**	**joy**	**fear**	**disgust**	**sadness**	**anger**	**surprise**	**contempt**
mean	13.75	0.01	1.76	0.29	0.65	1.07	0.18	10.37	0.02	8.49	0.20	1.01	0.66	0.18
median	0.00	0.00	0.39	0.01	0.00	0.22	0.19	0.00	0.00	0.43	0.02	0.00	0.20	0.19
std.dev	32.56	0.30	9.04	2.49	5.22	4.41	1.12	29.24	0.69	25.20	2.33	7.13	2.81	0.25

From the above analysis, we concluded that in Human-Computer Interaction, when the user is interacting with the avatar, the user is the *signaler* and the avatar shall behave like a human *builder*. The user's affective states may be very positive even when that of his/her partner are neutral, but their affects may also fluctuate immensely. Thus it is essential for the next generation of affective avatars to show human-like empathic facial expressions to help improving the user's affect and experience.

To investigate how *signalers* and *builders* affectively influence each other, we also conducted a descriptive analysis for the emotion metrics between *signalers* and *builders*. We first removed those lines in the spreadsheets that contained "nan" values because they were invalid data generated by Affdex [14]. Figure 3.8 and Figure 3.9 were the boxplots of seven basic emotion metrics of *builders* and *signalers*. The majority of data points were zeros thus resulted in a non-normal distribution for each emotion score sample. Those outliers were non-neutral emotions that spanned through 0 to 100 and included in our analysis because they contained important naturally elicited emotional features from the human subject.

Table 3.1 showed a summary table of the *signalers'* emotions and *builders'* emotions. In the table, we provided the mean, median and standard deviation of seven emotions (joy, fear, disgust, sadness, anger, surprise, and contempt). Very small scores were recorded for the emotions from fear to contempt. Indeed, joy was the most intense emotion in both *signalers* and *builders*, a high standard deviation meant it also had a spread distribution. The values of joy supported our hypothesis that overall *signalers* were happier than *builders* and different people experienced different levels of joy. Considering the significant presence of joy among all emotions, when we turned to recognize emotions from the user's face in Human-Computer interaction, we configured the perception module in Diana system to focus on recognizing joy. Disgust was the second strongest emotion in both *signalers* and *builders*, but due to disgust was not very possible to occur in our experimental context when two people faced each other, we omitted this emotion. Other emotions were very close to zero intensity, but in order to monitor the appearance of negative emotions in users, we also chose to configure the system to detect anger as a representation of the user's negative emotions.

3.3.2 Statistical Relationships between Human Affect

To analyze if there existed statistically significant difference in valence when a subject acted as a *builder* and a *signaler*, we ran a paired-sample sign test on every subject's paired valence scores. We selected the first trial of each human dyad settings, since the *signaler* and *builder* exchanged their roles at the midterm of experiment, 20 dyads resulted in 40 selected trials of tasks. In each trial, the valence scores from one CSV file had been averaged with the "nan" values excluded. Finally we obtained 20 pairs of valence scores in which the role of *signaler* or *builder* was treated as independent variables and the scores was the dependent variable.

We chose to use a paired-sample sign test because our data did not meet the normal distribution assumption of running a standard paired t-test, and the distribution of the differences between the two roles of one subject was not symmetrical in shape. Paired sign test is a less powerful non-parametric alternative to the paired t-test. Our null hypothesis H_0 was: p = Pr(S_v > B_v) = 0.5

meant that the valence scores of the *signaler* and the *builder* were equally likely to be greater than the other. H_A was: the valence scores of the *signaler* were greater than that of the *builder*. Let W be the number of valence pairs for which S_v - B_v > 0. Assuming that H_0 is true, then W follows a binomial distribution W $\sim$ b(m, 0.5). Computational result from the R software reported the role of *signaler* in a task elicited a statistically significant increase in valence scores compared to the role of *builder*, p = .003. The probability of a result that the valence of *signaler* greater than *builder* in 40 pairs was not likely to happen by chance, thus we concluded when a subject acts as the *signaler*, the valence scores were significantly greater than the scores when the subject acts as a *builder*, in other words, a person was generally happier when in the role of a *signaler* rather than a *builder*.

We then conducted statistical analysis on individual emotions between *signalers* and *builders*. The scores of emotion metrics were selected and averaged similarly as how we dealt with valence. For joy, the distribution was asymmetric, thus we ran a paired sign test again and got p = .001, indicated that the *signalers'* joy scores were greater than the *builders'* joy. This finding fitted our hypothesis that *signalers* had more positive emotions than *builders*. This might happened because *builders* were easier to feel stressed when they needed to interpret *signalers'* non-verbal instructions, while *signalers* could give any gestural instruction as they liked to finish the task. For negative emotions, we excluded the analysis of fear, disgust or surprise scores as they were not likely to occur during a Human-Human Interaction. In the remaining emotions, because the differences of anger, sadness, or contempt between *signalers* and *builders* had an approximately symmetric shape of distribution, we performed a Wilcoxon signed-rank test on these scores as it was more powerful than a paired sign test. A Wilcoxon signed-rank test is a non-parametric alternative of the paired t test, but it does not require the normality of paired data samples. It ranked the absolute difference between paired values then computed test statistics. Results indicated that there was no significant difference between *signaler* and *builders'* negative emotions anger (p = 0.38), sadness (p = 0.57), or contempt (p = 0.20).

Chapter 4

Modeling Empathy on Diana

This chapter presents two experiments. The first experiment concerned the calibration of Diana's facial expressions. In other words, how much of a facial expression on Diana's face was supposed to be. The second experiment investigated the user's evaluation of three modes of Diana: an emotionless Diana, a version of Diana that mimicked the user's facial expressions, and a Demo version of Diana that expressed dynamic facial expressions to model empathic behaviors.

4.1 Calibration of Diana's Facial Expressions

To design appropriate facial expressions, in the first experiment, we sent out questionnaires to the CS students at Colorado State University and collected 20 responses for each of the 4 facial expressions: joy, frustration, confusion, and concentration. In every question, Diana's facial expression from level A to E intensified from 20% to 100% of the scale of action units (i.e., the system scale of morph targets), such as the expression of joy showed in Figure 4.1. Figure 4.2 to Figure 4.5 were the pie charts of the distribution of users' votes we received. As for joy and frustration, more than 60% of participants agreed that the most intense facial expressions mostly expressed these two emotions. On the contrary, they preferred the least intense of facial expression to represent concentration. For confusion which was generally considered a high-level emotion, participants held different opinions that the people chose each level were nearly equally-distributed. Consider-

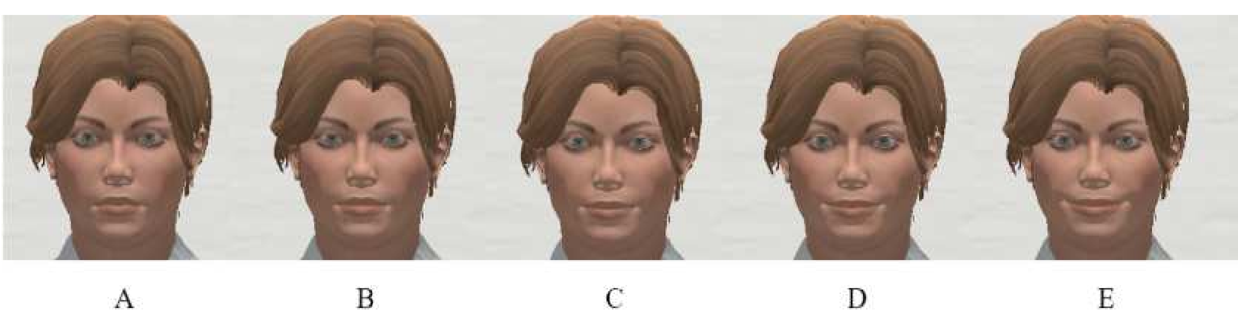

Figure 4.1: Diana's joy from level A to E intensified from 20% to 100% of the overall scale.

ing the analysis of our result, we set up Diana's joyful and sympathetic facial expressions to have nearly 100% of overall system scale, and weakened her confusion or concentration facial expressions to 20% of the system scale. These values worked as defined thresholds of action units in the following experiment.

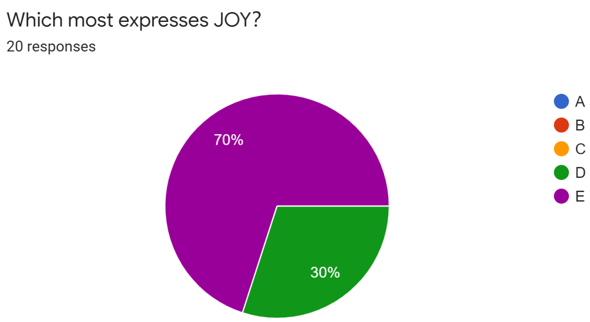

Figure 4.2: The pie chart of joy.

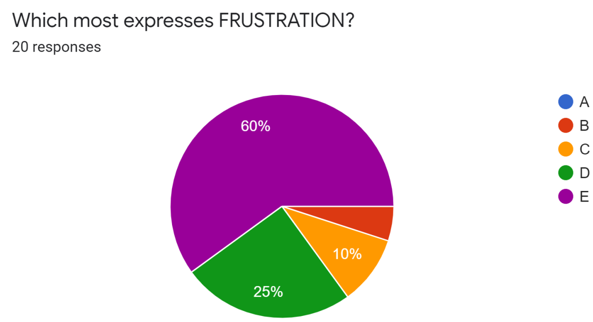

Figure 4.3: The pie chart of frustration.

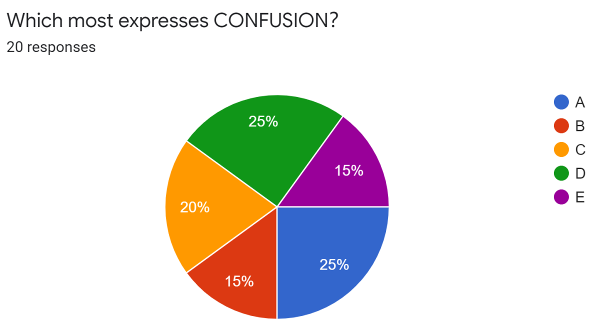

Figure 4.4: The pie chart of confusion.

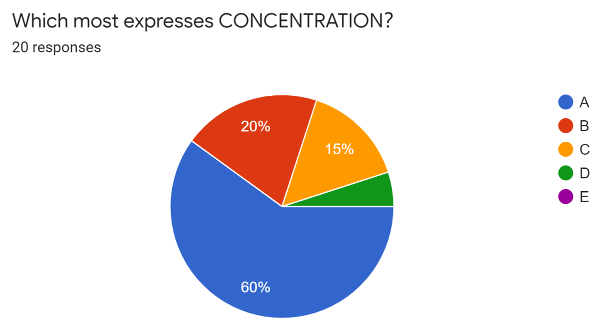

Figure 4.5: The pie chart of concentration.

4.2 Three Modes of Diana

The second experiment was an empirical human subject study. To give a comprehensive comparison of the user perception and performance between an emotionless avatar, an avatar that mimics the user's emotion, and an empathic avatar with dynamic affective states, we developed three modes of Diana in which the only difference was the update algorithm of her facial expressions. A brief description of the three modes of Diana is shown below:

- Emotionless: Diana maintained a flat facial expression throughout the experiment.
- Mimicry: Diana simultaneously expressed a joyful facial expression when the dominant emotion of the user was labeled joy. When the user's emotion was labeled angry, Diana showed a sympathetic facial expression that was modeled from sadness. If the user's emotion was then labeled as neutral, Diana immediately returned a neutral facial expression.
- Demo: Diana's responsive affect composed of a finite state machine [43] with five states: joy, sympathy, neutral, confusion, and concentration. With sympathy modeled from sadness and confusion modeled from our observation of human facial movements of action units, we aimed at providing consolation to the user by letting Diana act like a human builder. The states could transition between each other depending on the user's affect and gestures. Each state was entered when all the action unit values linearly moved and reached pre-defined thresholds, and each decay process took 2 seconds.

Specifically, in the Demo mode of Diana, different than the agent in the iViz Lab who expressed empathy verbally in three subsequent states: listening, thinking, and speaking, Diana was designed to react to the user's affects and gestures in terms of affective states. To make the occurrence of Diana's responsive affects perceived more like human facial expressions, we made the transitions of her affective states became smooth and linear. We constructed all her affective states into a form of a finite state machine. A finite state machine is a mathematical model of computation. It is an abstract machine that can be in exactly one of a finite number of states at any given time [43]. The states transitioned with respect to the conditions of **true** or **false** of the user's pointing status and the user's instant emotion labeled by the recognition module. A state was triggered or aggregated when a certain condition was met and Diana gradually intensified her facial expression until it reached a pre-defined threshold. If the user's gestural or emotional condition did not remain **true**, as time elapsed, every non-neutral facial expression gradually decayed.

A diagram of the finite state machine of Diana's affective states was shown in Figure 4.6. The orange lines represented conditions relevant to the user's affects and the blue lines represented

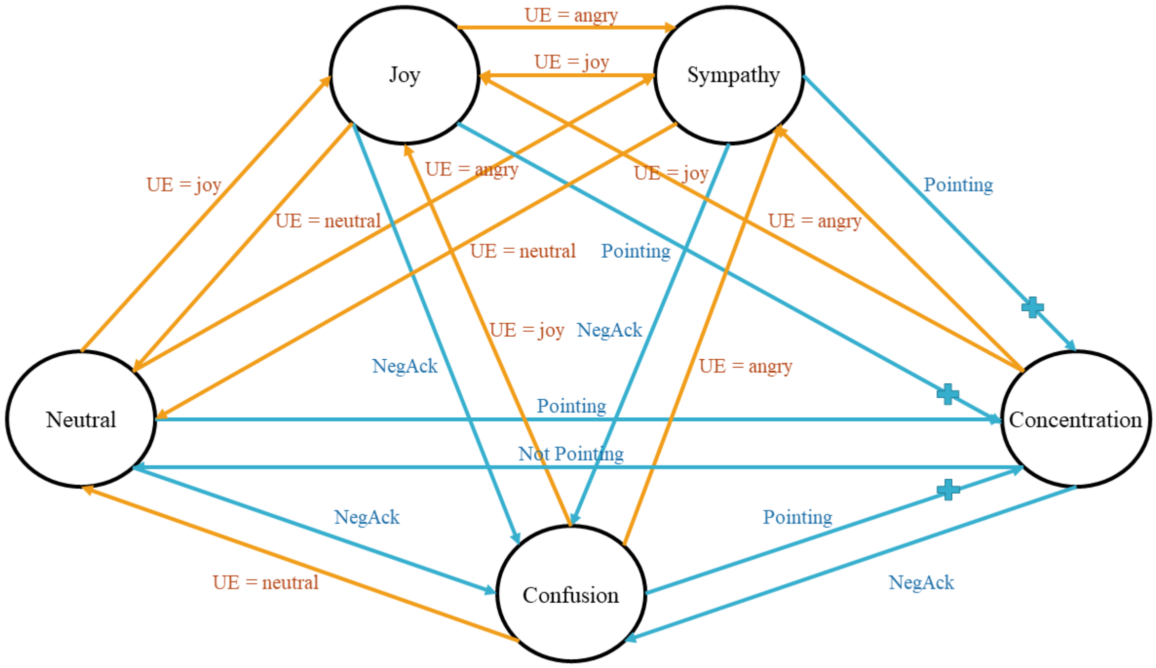

Figure 4.6: A diagram of the finite state machine of Diana's affective states.

conditions regarding the user's gestures (e.g., **true** or **false** of pointing status). UE indicated the user's instant emotion recognized and labeled by the recognition module, and NegAck meant a negative acknowledgment received from the user such as the never mind gesture.

The concrete conditions of transitions are:

- Diana greeted with a joyful facial expression when the user entered the interaction zone and started engaging.

- When the user was pointing, Diana's facial expression displayed concentration (e.g., with her eyes opened wider, brows higher). Emulating a human *builder* (from the EGGNOG dataset [2]), Diana provided a sense of patience.

- When Diana received a negative acknowledgment gesture (i.e., never mind) that indicated a mistake in her last action, or the user mentioned an object or showed a gesture that had not been defined, she displayed confusion (with a frown, etc).

- When Diana perceived that the user was happy, she showed a joyful facial expression. When she perceived the user was angry, she showed a sympathetic facial expression.

- If both the conditions of user gesturing and emotion were met, Diana's facial expressions were synthesized to form an aggregated state, which meant she could be joyful and concentrated, or sympathetic and concentrated at the same time.

The pseudo-code of the algorithm that updated Diana's responsive affect transitions between five states is shown below:

```
while FaceUpdating do
    Function NoteUserEngaged(userIsEngaged):
        // called once when the user starts/stops engaging
        if userIsEngaged then
            dianaEmotion = joy
        else
            dianaEmotion = neutral
    Function NoteUserBehavior(userEmotion, userIsPointing, dianaEmotion):
        // called when the user emotion or pointing status is changed
        if userEmotion = joy and userIsPointing then
            dianaEmotion = joy + concentration
        else if userEmotion = angry and userIsPointing then
            dianaEmotion = sympathy + concentration
        else if userEmotion = joy then
            dianaEmotion = joy
        else if userEmotion = angry then
            dianaEmotion = sympathy
        else if dianaEmotion != neutral and dianaEmotion != concentration then
            if userIsPointing then
                dianaEmotion gradually decays to concentration
            else
                dianaEmotion gradually decays to neutral
```

4.3 Design of Appropriate Responses

The second experiment carried out was designed to investigate how users responded to different forms of expression recognition and generation in Diana in the context of solving a task.

This experiment involved the Mimicry Diana and Demo Diana with responsive affect in terms of non-standard facial expressions because it was in the context of a collaborative task. As a comparison, many embodied conversational agents were designed to express Ekman's [1] seven basic emotions, see [15, 22, 23, 31, 32, 35]. However, these emotions are not all suitable to be expressed by an avatar in a collaborative environment. For instance, in such environment like ours where Diana and the user worked together to build blocks in a BlocksWorld, if the avatar expressed anger when the user was giving gestural instructions, an impatient user might get angry as well and the user performance might also be influenced.

Inspired by previous works and our findings in the data analysis of EGGNOG videos [2], we integrated four responsive affective states on Diana's face. Considering the difficulty of studies in the CS field to model empathy comprehensively [6], we tried to turn researchers previously proposed psychological concepts into software implementation, especially in terms of action unit combinations. When we designed Diana's facial expressions, the key concepts in her affect perception and generation modules were Thinking from Others Perspectives and the Appraisal Theory, they were components that resided in the highest level of the hierarchical model of empathy for embodied agents proposed by Yalcin et al. [6].

Table 4.1 showed Diana's action code combinations for expressions compared with the code combinations in standard Facial Action Coding System [1] and SmartBody [44]. SmartBody was a character animation platform originally developed at the USC Institute for Creative Technologies. SmartBody provided locomotion, steering, object manipulation, lip-syncing, gazing, nonverbal behavior and re-targeting in real-time. We summarized the action unit combinations from their agent's face animation and outlined them in the third column of the table. Regarding joy and sympathy, we developed our combinations based upon similar definitions in the Facial Action Coding System [1]. For confusion, we selected action units that were found to contribute to the perception of confusion. As for concentration, we proposed our creations by observing human behavior in EGGNOG [2]. All facial expressions also added with the action units that associated with high recognition accuracy and judgment of human-likeness [15]. Those missing action units in the character were replaced by movements of similar facial morph targets. At last, a synthesized facial expression was generated by linear movements towards pre-defined thresholds of the values of morph targets. The appearances of 4 non-neutral affective states are shown in Figure 4.7 to Figure 4.10.

Table 4.1: Diana's action unit code combinations compared with other code combinations.

Affective States	FACS [1]	SmartBody [44]	Diana's Action Units
Joy	6 + 12 (Happiness)	same	BrowsUp + NoseScrunch + MouthNarrow + Smile
Sympathy	1 + 4 + 15 (Sadness)	1 + 4 + 6	BrowsOuterLower + BrowsDown + Frown + NoseScrunch + MouthNarrow
Confusion	4 + 7 + 15 + 17 + 23 [45]	-	BrowsIn + Squint + NoseScrunch + JawDown
Concentration	-	-	BrowsUp + EyesWide

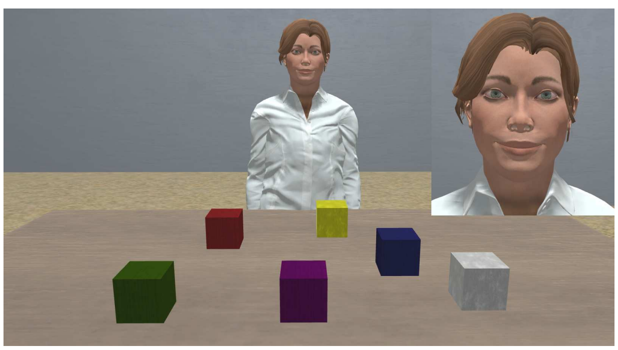

Figure 4.7: The affective state of joy.

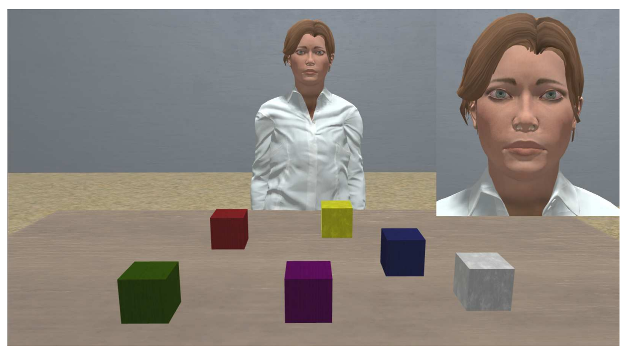

Figure 4.8: The affective state of sympathy.

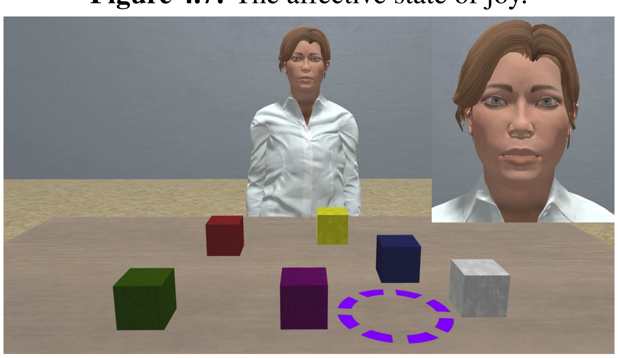

Figure 4.9: The affective state of confusion.

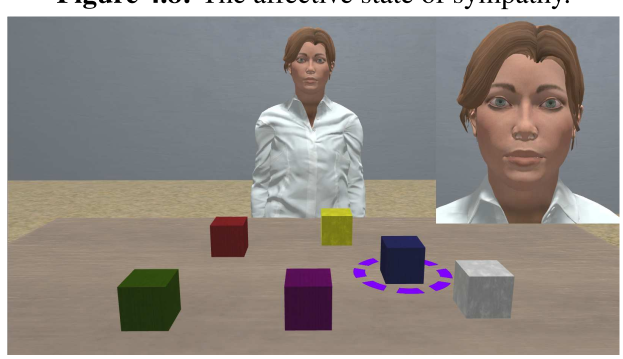

Figure 4.10: The affective state of concentration.

4.4 Experimental Setup

In this section, we talk about the participants' demographic information, the experimental equipment being used, and the procedure of carrying out experiment and evaluating the three modes of Diana.

4.4.1 Participants

The experiment consisted of 21 participants (9 female and 12 male) with ages between 19 to 46 (M = 25.33, SD = 7.47). One additional subject was not able to finish all the tasks, therefore, her data was not included in the following analysis. These volunteers were recruited through emails and word of mouth. Subjects were composed of undergraduate, graduate students (mostly in CS major), and staff at Colorado State University. As reported in the demographic questionnaire, four subjects had experience with virtual agents/avatars before. Eighteen subjects had prior experience of playing games using gaming devices (e.g., Nintendo Switch, Kinect [11]), within these people, six used to play games a lot but in recent years they had shortened the time to 1-2 hours a week.

4.4.2 Equipment

The experiment was conducted with Diana system running in a laptop that projected on a desktop monitor for display. Diana system was developed in Unity Editor 2018.4.16f1. The laptop ran a Windows 10 professional system with an Intel i9-9900k 3.6GHz processor and an NVIDIA GeForce RTX 2080 graphics card. The users' interactions with Diana were recorded by using the OBS studio application, the RGB-D data was captured by the Microsoft Kinect v2 sensor [11] and the RGB frames for emotion recognition were captured by an HP 4310 webcam. Considering if the users spoke, they might not express as many affects as they were concentrating on observing Diana, we did not allow the usage of verbal signals in our experiment, thus the Yeti microphone was muted.

4.4.3 Procedure

For consistency, only one researcher presented and conducted the experiment in the lab. At the start of each session, participants were asked to fill in a consent form and a video release form, the forms claimed our right of using collected data for research purposes. Before participants came to the lab, they had also finished an online demographic questionnaire investigating their identity, the amount of facial hair, and game usage, etc. Then the participants were shown a 3-minute video introducing the procedure of the experiment and the gestures they were allowed to use. Subjects were told to only use gestures as instructions. To simplify the task, gestures included in experiment were: waving, pointing, and never mind. The goal of the task was also revealed as moving the 6 different color blocks on the table to form a horizontal straight line with blocks next to each other (the color order was assigned as: red, orange, yellow, green, blue, and purple). Participants were then given enough time to ask questions, practice with those gestures, and move blocks to form any structure as they like. The measures of experiment only began when participants told the researcher they were familiar with the gestures and ready for the task.

The experiment followed a within-subject (i.e., repeated-measures) design. The independent variable was the mode of avatar: Emotionless, Mimicry, and Demo. The order of modes in interactions was not revealed to users during the experiment, but to help users distinguish between avatars, the backwall color of Emotionless Diana, Mimicry Diana, and Demo Diana were set to be white, light green, and light blue, respectively. The three modes of Diana followed a permuted block randomization which was a way to randomly allocate a participant to a treatment group while maintaining a balance across treatment groups [46]. Each "block" had one randomly ordered treatment assignment of the modes. The order from permutation associated with the subject ID was output into a text file. For each mode (e.g., Emotionless Diana), the subject repeated the same task with her three times. For accurate recording, a logger was built in the system to record the timestamps and events that happened in each trial.

When one trial of one mode began, the researcher typed in the subject ID and started the virtual scene and the recording manually. The first mode of Diana in a BlocksWorld associated with that

ID in the text file was read and displayed on the screen. Then the participant walked into the interaction zone and started interacting. The original location of every block on the table was fixed for each trial. Once the key "user:isEngaged" became **true**, the logger formatted the subject ID, the mode of avatar, a string of event "Start engaging", followed by the timestamp of date (yyyy-MM-dd) and hour (HH:mm:ss.SSS). Subsequently, once the y coordinate of all six blocks in the BlocksWorld were equal, the blocks were considered as forming a horizontal straight line, and the logger formatted the same attributes again and replaced the event string with "Finish Task" this time. The subject was asked to quit the interaction zone after they finished every trial, and the researcher then pressed a key on the keyboard to replay this scene. This process was repeated until the subject finished three trials with this mode of Diana. The video recording was then paused and the subject filled in a 5-point Likert scale questionnaire and a NASA Task Load Index survey [47]. The researcher then pressed another key on the keyboard to switch to the next mode of Diana. Then the subject started interaction again and the recording was resumed. The questionnaire was the same and was given to the subject after the third trial with each mode of Diana, except at the last of the experiment the subject was asked to fill in which mode of Diana was their favorite and was required to give the reason.

The first questionnaire provided to the user after interaction was a 5-point Likert scale questionnaire asking about their perceptions and experience about the avatar they just interacted with, (with Strongly Disagree = 1; Disagree = 2; Neutral = 3; Agree = 4; Strongly Agree = 5). At the end of each questionnaire, there was also a optional text box provided for subjects to type in comments. In this questionnaire, we asked 7 unbiased questions investigating users' perceptions of the whole character's movements and appearance including 3 questions regarding the avatar's facial expressions. We did so to prevent giving focused questions that might imply the users to give positive responses of Diana's facial expressions. In each mode of Diana, the Cronbach's α of subjects' answers achieved 0.845, 0.836, and 0.8 respectively, indicated that all the statements in the questionnaire followed a good internal consistency.

4.5 Results

4.5.1 Positive Responses

Table 4.2: Raw count of positive responses in three modes of Diana regarding each question.

Question Number	Question Wording	Avatar Mode Discription	Avatar Mode Label	Positive Responses
1	The avatar looks friendly	Empathic Affect	D	11
		No Affect	E	10
		Mimic User's Affect	M	10
2	It helped to look at the avatar's face	Empathic Affect	D	4
		No Affect	E	3
		Mimic User's Affect	M	3
3	The avatar was helping me	No Affect	E	10
		Mimic User's Affect	M	9
		Empathic Affect	D	7
4	The avatar's facial expressions are natural	Empathic Affect	D	8
		No Affect	E	7
		Mimic User's Affect	M	6
5	The avatar's movements are natural	Empathic Affect	D	10
		Mimic User's Affect	M	7
		No Affect	E	6
6	I felt comfortable working with this avatar	Empathic Affect	D	13
		No Affect	E	12
		Mimic User's Affect	M	11
7	I felt relaxed working with this avatar	Empathic Affect	D	10
		Mimic User's Affect	M	10
		No Affect	E	9

The raw count of positive responses each mode of Diana received on each question were shown in Table 4.2. The table provided question statements, descriptions of avatar modes, labels of avatar modes, in which "D" meant Demo, "E" represented Emotionless, and "M" referred to Mimicry, followed by the raw count of positive responses corresponded to the sum of votes participants gave for answers "Agree" and "Strongly Agree". In 5 out of 7 questions, the Demo mode of Diana received more positive votes than the other two modes, while in the last question it received equal number of positive responses as the raw counts for mode Mimicry.

4.5.2 Distribution of Ratings Votes

Figure 4.11 showed a diverging stacked bar chart that plotted the number of users votes in percentages. Every three adjacent bars corresponded to an individual question in the questionnaire, and from top to bottom the three bars represented results of mode Mimicry, Emotionless, and Demo, respectively. To simplify comparison of distributions, all horizontal bars were aligned at a vertical dividing line that separated Neutral and Agree, with the percents from Strongly Disagree to Neutral were marked as negative values and the percents from Agree to Strongly Agree were marked as positive values.

While other questions had an approximately half-half distribution of the positive and negative votes, the question "It helped to look at the avatar's face" received much more negative responses than positive responses in all three modes. This might be caused by the condition that when users were giving gestural instructions to the avatar, they put their concentration mainly on the blocks movement and Diana's arm motion instead of her face, and thus they did not consider looking at Diana's face was helpful (as one participant commented). Besides that, in the question "I felt relaxed working with this avatar", users gave almost the same number of positive responses for each mode, and in the question "The avatar was helping me", mode Demo received the least positive responses, we attributed this to the phenomenon in previous findings that some subjects reported they felt anxious when they were being "watched" by an embodied agent. The deeper reason needs to be further investigated. However, mode Demo beat the other two modes by having gained more positive responses in the rest of the questions.

Because in the three modes of Diana not every rating score sample was normally distributed, we chose to conduct Friedman's test as a non-parametric alternative to the one-way repeated-measure ANOVA. Friedman's test was used on a matrix with n rows (blocks), k columns (treatments), and there was only one observation at the intersection of each block and treatment. In our study, the block corresponded to the ID of the participant and the treatment was the mode of Diana. The ranks within each block were calculated and the test statistic was computed. We compared the medians of the numerical ratings 1-5 in each individual question for each mode of Diana. Results

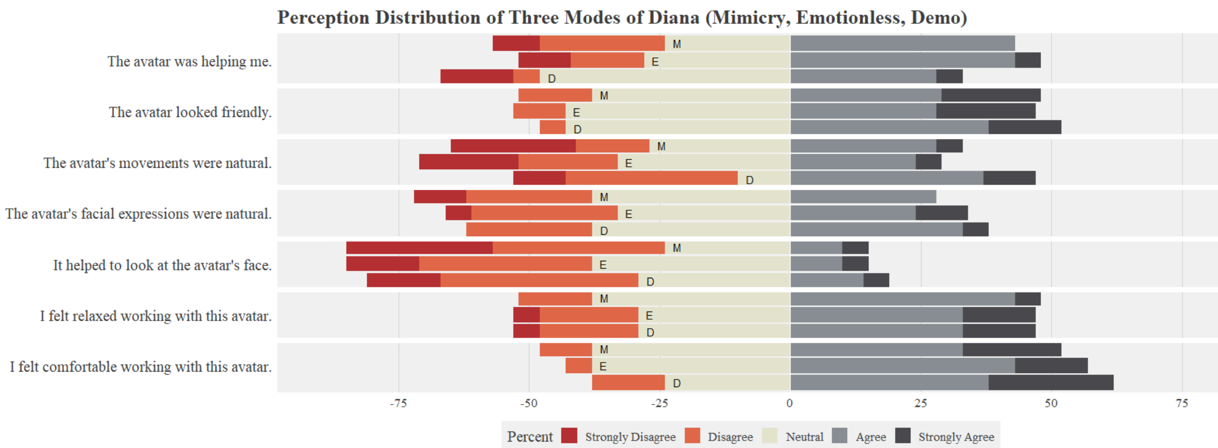

Figure 4.11: Diverging Stacked Bar Chart of the percentages of users' perceptive votes.

showed no significant difference between the votes in three modes of Diana, as the statistics and *p* values shown in Table 4.3.

Table 4.3: Friedman's test results for individual Likert scale questions.

Question	χ^2	DoF	p-value
The avatar looked friendly.	0.13	2	0.94
It helped to look at the avatar's face.	4.92	2	0.09
The avatar was helping me.	1.24	2	0.54
The avatar's facial expressions were natural.	2.46	2	0.29
The avatar's movements were natural.	4	2	0.14
I felt comfortable working with this avatar.	0.84	2	0.66
I felt relaxed working with this avatar.	0.27	2	0.87

After finished the last trial of the experiment, when we asked about which mode of Diana was preferred, 21 participants gave 9 votes to the Mimicry avatar, with 7 votes given to the Demo avatar and the remaining 5 votes to the Emotionless avatar. Users also subjectively commented the mode of avatar that had natural facial expressions and made them feel comfortable to work with. The result suggested that the two affective avatars were perceived by users as more natural and friendly to interact with compared to the Emotionless avatar. But 3 participants also mentioned the Uncanny Valley effect [38] and reported the avatar's facial expressions made them think they did something wrong.

4.5.3 NASA Task Load Index

The NASA Task Load Index is a widely used, subjective, multidimensional assessment tool that rates perceived workload in order to assess a task [16]. In our experiment, we used it as a survey to measure 6 metrics (mental demand, physical demand, temporal demand, performance, effort, and frustration) of the task. Table 4.4 showed a descriptive analysis of averaged scores in a summary table. After checked assumptions, a Friedman's test was conducted on the results of NASA TLX individual questions and indicated there was no significant difference between the NASA TLX scores in three modes of Diana, as shown in Table 4.5. It meant the perceived work load of three modes of Diana were the same.

Table 4.4: Averaged NASA TLX scores by mode.

	Demo	Mimicry	Emotionless
Mean	33.05556	33.45238	32.77778
SD	16.96265	15.08376	18.49800

Table 4.5: Friedman's test results for individual NASA TLX metrics.

Metric	χ^2	DoF	p-value
Mental Demand	0.08	2	0.96
Physical Demand	0.19	2	0.91
Temporal Demand	0.43	2	0.81
Performance	1.34	2	0.51
Effort	1.34	2	0.51
Frustration	0.86	2	0.65

4.5.4 Length of Completion

We conducted an objective measurement by recording the length of task completion for every subject when they interacted with each mode. The logger was invoked when the user stepped into the interaction zone and stopped when all the blocks on the table formed a horizontal line. Again Friedman's test was conducted but there was no significant difference in completion time in the three modes of Diana. Descriptive analysis showed mode Demo (*M* = 73.90, *SD* = 20.19) took slightly longer time than mode Emotionless (*M* = 70.38, *SD* = 23.67) and mode Mimicry (*M* = 65.90, *SD* = 20.12).

The trial completion time in seconds as measured between a participant entered the interaction zone and all the blocks formed a horizontal straight line are shown in Table 4.6. From observation, mode Demo seemed took the longest time in completing a task, followed by emotionless and mimicry. Because the residuals in the trial time of each mode were not normally distributed, we ran the Friedman's test again. There was no significant difference between trial completion times in three modes of Diana ($\chi^2(2) = 1.61$, $p = 0.45$).

Table 4.6: Average trial completion time by mode in seconds.

	Demo	Mimicry	Emotionless
Mean	73.90476	65.90476	70.38095
SD	20.19135	20.12189	23.67377

4.6 Discussion

In summary, we first found the difference in affective metrics between *signalers* and *builders* when they worked together on a task in the BlocksWorld. Our results suggested overall the *signalers'* affects were more broad in the valence scale than *builders*, indicating they experienced more varied emotions. We also found joy was the most commonly elicited emotion than any other basic emotion during *signaler-builder* interactions, thus marked joy as the most important emotion to observe during Human-Computer Interaction. In summary, *signalers* were generally happier than *builders*. For individuals, when one subject acted as the *signaler*, he/she was happier than working as a builder.

The results from our empirical human subject study indicated participants perceived relatively intense joyful facial expressions, and they perceived different intensities of more complex states such as frustration or concentration. We obtained very similar perceptions and experience scores in three modes of Diana. Though not statistically significant, the Demo mode received a few more positive votes in the perception questionnaire and took users more seconds to finish a task. The insignificant result might be caused by confounding variables in our experimental setup. First, the whole study was conducted under a context of building blocks with hand gestures and observing Diana's behaviors, during the task, participants might forget to look at Diana's face and instead focused on her arm motions and blocks if not reminded by the researcher. Second, the dynamic and smooth transitions of Diana's affective states added more difficulty to observe the changes on Diana's face. Participants need to be looking at her face at the appropriate time point between transitions and kept observing until a new expression was fully presented so that they would notice the difference, which was not likely to happen in real interactions as users could get distracted. Third, in the two modes Demo and Mimicry that with affective ability, the appearance of Diana's affective states were largely depending on the participant's affects. Only a relatively emotional user would elicit an emotive avatar. In other words, if the user kept a neutral face or did not use the never mind gesture throughout the experiment, the difference between the three modes of Diana would be very subtle and difficult to find out.

In the last section of the questionnaire, participants also left many useful comments and suggestions for Diana. Some participants had noticed the different facial expressions between the three modes of Diana. Due to we asked about their perceptions regarding Diana's overall movement, they also provided comments on Diana's behaviors of arm motion. The Uncanny Valley effect still existed in our experiment, as three participants left comments for the Mimicry mode of Diana like "She seemed not quite as creepy as the third one (Demo) because I could feel less from her.", "This felt the most natural to me. The other two facial reactions and movements felt quite strange to me.", "Her facial design was slightly In the Uncanny Valley." However, the Demo mode of Diana also received positive comments like "This avatar has more facial expressions to me.", "This one is more natural because it smiled and the suitable body's movement.", "The facial expression and movement of the avatar seem the most natural to me.", "This one looked more friendly." There was also one negative perception elicited: "I felt the avatar was angry with me when I did something wrong." These comments showed a generally positive attitude on the affective Diana and her empathic facial expressions.

4.7 Limitation of the Study

There were also some limitations of our study: First, in the Unity platform, we could only control a limited set of action units, that introduced difficulty when synthesizing some complex facial expressions because we could only either omit or replace some standard action units, thus the final effect might not be as expressive as the natural facial expressions defined in the Facial Action Coding System [1]. Second, although we tried our best to mitigate the Uncanny Valley effect by adjusting the intensity of facial expressions on Diana's face, some participants still commented they experienced uncomfortable perceptions when interacted with the avatars that showed facial expressions. This situation might get improved by refining the texture of the character or designing more fine-grained facial expressions.

Besides the software restrictions, there was also an experimental limitation. In the previous elicitation study of EGGNOG [2], the tasks were randomly selected from a layout set and assigned to the human dyads, the number of blocks used and the structure pattern were all different, resulted in various length of completion of tasks and introduced confounding variables like levels of difficulty that might impact *signaler* and *builder* affect. Considering the time restriction in our repeated-measure study design, we assigned all subjects the same simple task of building a horizontal straight line. All of the participants could finish the task once in two minutes, this was a relatively short session of interaction compared to other human subject studies. This setting might be too short for either the user or Diana to fully elicit and express their facial expressions. In the future, we could create a more immersive and interactive experience for users such as an assembly task of toys to better recognize and interpret human affect.

Chapter 5

Conclusion and Future Work

In our study, we proposed an affective avatar whose behavior was upon that of a person adopting the task role later taken up by the avatar. To investigate the role of affect when users and avatar jointly solving a task, we carried out a study to test if adding human affect to a collaborative task-focused avatar would improve the user experience and result in faster task completion. As affect has been massively studied in the field of Affective Computing because of its complexity across fields, we gave our avatar emotional intelligence to be able to recognize, interpret, and simulate human affect especially empathy. To model human affective states, our avatar not only expressed basic emotions such as joy, but also showed more complex affective states like confusion, concentration, and sympathy in terms of facial expressions based upon observations of human affective behaviors.

We also carried out an analysis of over 7 hours of video involving pairs of people solving blocks world tasks such as those solved by a person and the avatar. The footage of human-human interactions composed our EGGNOG dataset [2] that aimed at eliciting naturally occurred behaviors from naive users. This dataset is also a large-scale video resource for designing natural human-computer interfaces. When we analyzing affect, we focused on comparing the valence (the positive or negative nature of emotions) and seven basic emotions between the *signalers* and *builders*. Joy was the most significant emotion among pairs of people during collaborations, it also occurred with different levels of valences on different people. In those studies, it was found that the person who gave instructions (the *signaler*) was more emotional and generally happier than the person who interpreted signals and moved blocks accordingly (the *builder*). We also found out the level of joy was more associated with the role in a task rather than an individual person. When a pair of subjects was asked to exchange roles in finishing tasks, the subjects were happier to act as *signalers* compared to the cases when they acted as *builders*. This situation makes sense because when finishing a task, the *builder* who needs to interpret the *signaler's* gesture and move blocks are always more stressed. A *builder* may worry about if he/she has moved the block to the correct position that satisfied the *signaler*, while the *signaler* may try to act relaxed and happily to encourage the *builder*.

Besides joy, other basic emotions including fear, disgust, sadness, anger, surprise, and contempt were all weak during collaborations. there was no difference in other basic emotions between *signalers* and *builders*. At last, considering the possibility of occurrence, we chose joy and anger as two representative user emotions to be configured to recognize in Diana's perception module.

We utilized those findings on human affect as guidance to design our affective avatar called Diana as a human-like *builder*. To cooperate with a potential emotional *signaler*, Diana's affective states were designed to be expressed in terms of facial expressions, and Diana showed her empathy by thinking from the user's perspective, i.e., showing concentration when the user was pointing. The facial expressions were composed of linear combinations of the morph targets of action units that were defined in the Facial Action Coding System [1], along with individual action units that could improve recognition accuracy or human-likeness. A pre-study survey was sent out to investigate how strong did users suppose her facial expressions to be. For joy and frustration (which was designed to express sympathy in later studies), users preferred the most intense facial expressions, but for more complex emotions confusion and concentration, the number of users chose different levels of facial expressions were nearly equal and preferred the weakest expression, respectively. Thus we intensified the expression of joy state and weakened the sympathy, confusion, and concentration states on Diana's face.

We also added a dynamic architecture to Diana's affective states. The five states: neutral, joy, sympathy, confusion, concentration were fully connected and could transition between each other depending on user emotions and gestures. The transitions were linear and slow decreases and increases of the intensities of facial expressions. These features made our avatar's facial movements perceived by users as more natural and smooth just like human *builders*.

An empirical human subject study was conducted between a Demo mode of Diana with dynamic affective states, a Mimicry mode of Diana who mimicked users' instant facial expressions, and an Emotionless Diana with a flat face. Twenty-one subjects interacted with all three modes of Diana in a repeated design experiment. Objective measurement included the time of completion of each task, subjective measurements included the rating scores in a 5-point Likert scale post-study questionnaire and a NASA TLX survey. Questions in the Likert scale questionnaire asked about users' feelings about Diana's facial expressions and movements. The NASA TLX measured the task load by asking the users questions regarding mental demand and physical demand, etc. Though the scores of user perception between three modes of Diana in these two questionnaires were not statistically significantly different, participants spent a little more time with the Demo Diana, followed by the Mimicry Diana and the Emotionless Diana. They also gave the Demo one more positive votes in 5 out of 7 questions. In the comments for three modes of Diana, users could perceive the facial expressions of the Mimicry and Demo avatar and said they were natural. Results indicated that adding affective states on Diana led to a longer time for the user to finish the task. This may be caused by users spent more time observing Diana's facial expressions. Though some users were elicited feelings related to the Uncanny Valley effect, some other users rated the Demo Diana as friendly and comfortable to collaborate with, which meant our emphatic avatar was considered as a reliable partner in Human-Computer Interaction.

Our research added to previous works another quantitative analysis on human affect differences especially in collaborative contexts, and also provided evidence on user preference of an affective avatar rather than an emotionless avatar. Regarding the synthesized facial expressions that worked as a combination of previous findings and our own creations to represent non-basic emotions, participants could perceive it as natural facial expressions, indicating our method was another practical way of generating human-like facial expressions that express more complex human affect such as sympathy, concentration, and confusion. Our work presented another step in designing natural 3D user interfaces.

In the future, we plan to conduct more experiments on the collaboration between humans and an affective avatar. There were still unresolved problems in this thesis because although users could recognize the facial expressions on Diana's face, users did not perceive the Demo avatar and the Mimicry avatar very differently, indicating the shortage of a complete implementation in designing a high-level empathic avatar. Both positive and negative votes indicated our approach of adding empathic behaviors was still risky. By modeling dynamic facial expressions using deep neural networks, the user perception of Diana's facial expressions may get improved. Before that, let the avatar only mimic the user's facial expressions may be a safer choice.

It is clear that human affect is still a complicated signal in Human-Computer Interaction. The recognition of human affect and the generation of avatar's facial expressions are the very beginning technical steps in research, next steps include an autonomous affective agent that interprets human affect like a real human when collaborating with users. Overall, to create a real empathic avatar in natural 3D user interfaces, we shall keep the method of designing human-centered machines so the avatar's affect shall be modeled from real empathic human affect.

Bibliography

[1] Paul Ekman, Wallace V Friesen, Joseph C Hager, and A Human Face (Firm). Facial Action Coding System. *Salt Lake City, UT: A Human Face*, 2002.

[2] Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J Ross Beveridge, Bruce A Draper, and Jaime Ruiz. Eggnog: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 414–421, 2017.

[3] Appraisal Theory. https://en.wikipedia.org/wiki/Appraisal_theory.

[4] Affective Computing. https://en.wikipedia.org/wiki/Affective_computing.

[5] Raoul Rickenberg and Byron Reeves. The effects of animated characters on anxiety, task performance, and evaluations of user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 49–56, 2000.

[6] Özge Nilay Yalcin and Steve DiPaola. A computational model of empathy for interactive agents. *Biologically inspired cognitive architectures*, 26:20–25, 2018.

[7] Affectiva Human Perception AI Analyzes Complex Human States. https://www.affectiva.com.

[8] Siri-Apple. https://www.apple.com/siri.

[9] Alexa. https://www.alexa.com.

[10] Gururaj Mulay, Bruce A Draper, and J Ross Beveridge. Adapting rgb pose estimation to new domains. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0324–0330. IEEE, 2019.

[11] Kinect for windows. https://developer.microsoft.com/en-us/windows/kinect.

[12] Speech-to-Text accurately convert speech into text using an api powered by google's ai technologies. https://cloud.google.com/speech-to-text.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: a cross-platform real-time multi-face expression recognition toolkit. *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 3723–3726, 2016.

[15] Chaona Chen, Laura B Hensel, Yaocong Duan, Robin AA Ince, Oliver GB Garrod, Jonas Beskow, Rachael E Jack, and Philippe G Schyns. Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[16] Nasa-Tlx. https://en.wikipedia.org/wiki/NASA-TLX.

[17] Erwin Marsi and Ferdi Van Rooden. Expressing uncertainty with a talking head in a multimodal question-answering system. In *MOG 2007 Workshop on Multimodal Output Generation*, page 105, 2007.

[18] Susanne Van Mulken, Elisabeth André, and Jochen Müller. The persona effect: How substantial is it? In *People and computers XIII*, pages 53–66. Springer, 1998.

[19] Amy L Baylor and Yanghee Kim. Simulating instructional roles through pedagogical agents. *International Journal of Artificial Intelligence in Education*, 15(2):95–115, 2005.

[20] Michael Kipp, Kerstin H Kipp, Alassane Ndiaye, and Patrick Gebhard. Evaluating the tangible interface and virtual characters in the interactive cohibit exhibit. In *International Workshop on Intelligent Virtual Agents*, pages 434–444. Springer, 2006.

[21] Lisa Fan, Matthias Scheutz, Monika Lohani, Marissa McCoy, and Charlene Stokes. Do we need emotionally intelligent artificial agents? first results of human perceptions of emotional intelligence in humans compared to robots. In *International Conference on Intelligent Virtual Agents*, pages 129–141. Springer, 2017.

[22] Samuel Marcos Pablos, Jaime Gómez García-Bermejo, Eduardo Zalama Casanova, and Joaquín López. Dynamic facial emotion recognition oriented to hci applications. *Interacting with Computers*, 27(2):99–119, 2013.

[23] Gwo-Dong Chen, Jih-Hsien Lee, Chin-Yeh Wang, Po-Yao Chao, Liang-Yi Li, and Tzung-Yi Lee. An empathic avatar in a computer-aided learning program to encourage and persuade learners. *Journal of Educational Technology & Society*, 15(2):62–72, 2012.

[24] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S Huang. Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007.

[25] Laurence Devillers, Ioana Vasilescu, and Lori Lamel. Annotation and detection of emotion in a task-oriented human-human dialog corpus. *proceedings of ISLE Workshop*, 2002.

[26] Laurence Devillers, Lori Lamel, and Ioana Vasilescu. Emotion detection in task-oriented spoken dialogues. *2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698)*, 3:III–549, 2003.

[27] Laurence Devillers, Ioana Vasilescu, and Catherine Mathon. Prosodic cues for perceptual emotion detection in task-oriented human-human corpus. *15th International Congress of Phonetic Sciences*, 2003.

[28] Saif M Mohammad and Felipe Bravo-Marquez. Wassa-2017 shared task on emotion intensity. *arXiv preprint arXiv:1708.03700*, 2017.

[29] Best-worst_scaling. https://en.wikipedia.org/wiki/Best-worst_scaling.

[30] Aurélie Zara, Valérie Maffiolo, Jean Claude Martin, and Laurence Devillers. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 464–475. Springer, 2007.

[31] Jeonghun Ku, Hee Jeong Jang, Kwang Uk Kim, Jae Hun Kim, Sung Hyouk Park, Jang Han Lee, Jae Jin Kim, In Y Kim, and Sun I Kim. Experimental results of affective valence and arousal to avatar's facial expressions. *CyberPsychology & Behavior*, 8(5):493–503, 2005.

[32] Jie Shen, Ognjen Rudovic, Shiyang Cheng, and Maja Pantic. Sentiment apprehension in human-robot interaction with nao. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 867–872, 2015.

[33] Deepali Aneja, Daniel McDuff, and Shital Shah. A high-fidelity open embodied avatar with lip syncing and expression capabilities. In *2019 International Conference on Multimodal Interaction*, pages 69–73, 2019.

[34] Nicholas V Mudrick, Michelle Taub, Roger Azevedo, Jonathan Rowe, and James Lester. Toward affect-sensitive virtual human tutors: The influence of facial expressions on learning and emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 184–189. IEEE, 2017.

[35] Pablo Gomez Esteban and David Rios Insua. An affective model for a non-expensive utility-based decision agent. *IEEE Transactions on Affective Computing*, 2017.

[36] Jeffrey F Cohn. Foundations of human computing: facial expression and emotion. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 233–238, 2006.

[37] Kerstin Haring, Kristin Nye, Ryan Darby, Elizabeth Phillips, Ewart de Visser, and Chad Tossell. I'm not playing anymore! a study comparing perceptions of robot and human cheating behavior. In *International Conference on Social Robotics*, pages 410–419. Springer, 2019.

[38] Karl F MacDorman and Hiroshi Ishiguro. The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies*, 7(3):297–337, 2006.

[39] Mouna Belhaj, Fahem Kebair, and Lamjed Ben Said. Emotional dynamics and coping mechanisms to generate human-like agent behaviors. *Applied Artificial Intelligence*, 31(5-6):472–492, 2017.

[40] Klaus R Scherer, Marcello Mortillaro, Irene Rotondi, Ilaria Sergi, and Stéphanie Trznadel. Appraisal-driven facial actions as building blocks for emotion inference. *Journal of personality and social psychology*, 114(3):358, 2018.

[41] Luis-Felipe Rodríguez and Félix Ramos. Computational models of emotions for autonomous agents: major challenges. *Artificial Intelligence Review*, 43(3):437–465, 2015.

[42] Pradyumna Narayana, Nikhil Krishnaswamy, Isaac Wang, Rahul Bangar, Dhruva Patil, Gururaj Mulay, Kyeongmin Rim, Ross Beveridge, Jaime Ruiz, James Pustejovsky, et al. Cooperating with avatars through gesture, language and action. In *Proceedings of SAI Intelligent Systems Conference*, pages 272–293. Springer, 2018.

[43] Finite-state machine. https://en.wikipedia.org/wiki/Finite-state_machine.

[44] SmartBody. https://smartbody.ict.usc.edu.

[45] Tools for the facial action coding system (facs). https://www.noldus.com/applications/facial-action-coding-system.

[46] Permuted-block-randomization. https://www.statisticshowto.com/permuted-block-randomization.

[47] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

Appendix A

License

Colorado State University LaTeX Thesis Template

by Elliott Forney – 2017