Multimodal Technologies and Interaction

MDPI

Article

Building an Emotionally Responsive Avatar with Dynamic Facial Expressions in Human—Computer Interactions

Heting Wang [1,†], Vidya Gaddy [2,*,†], James Ross Beveridge [2,†] and Francisco R. Ortega [2,*,†]

[1] Computer Science Department, University of Florida, Gainesville, FL 32611, USA; heting.wang@ufl.edu
[2] Computer Science Department, Colorado State University, Fort Collins, CO 80521, USA; Ross.Beveridge@colostate.edu
* Correspondence: gaddvi@colostate.edu (V.G.); fortega@colostate.edu (F.R.O.)
† These authors contributed equally to this work.

Abstract: The role of affect has been long studied in human–computer interactions. Unlike previous studies that focused on seven basic emotions, an avatar named Diana was introduced who expresses a higher level of emotional intelligence. To adapt to the users various affects during interaction, Diana simulates emotions with dynamic facial expressions. When two people collaborated to build blocks, their affects were recognized and labeled using the Affdex SDK and a descriptive analysis was provided. When participants turned to collaborate with Diana, their subjective responses were collected and the length of completion was recorded. Three modes of Diana were involved: a flat-faced Diana, a Diana that used mimicry facial expressions, and a Diana that used emotionally responsive facial expressions. Twenty-one responses were collected through a five-point Likert scale questionnaire and the NASA TLX. Results from questionnaires were not statistically different. However, the emotionally responsive Diana obtained more positive responses, and people spent the longest time with the mimicry Diana. In post-study comments, most participants perceived facial expressions on Diana's face as natural, four mentioned uncomfortable feelings caused by the Uncanny Valley effect.

Keywords: human–computer interaction; affective computing; facial expression

check for updates

Citation: Wang, H.; Gaddy, V.; Beveridge, J.R.; Ortega, F.R. Building an Emotionally Responsive Avatar with Dynamic Facial Expressions in Human—Computer Interactions. *Multimodal Technol. Interact.* **2021**, *5*, 13. https://doi.org/10.3390/mti5030013

Academic Editors: Kangsoo Kim, Sun Joo (Grace) Ahn and Gerd Bruder

Received: 29 December 2020
Accepted: 16 March 2021
Published: 20 March 2021

1. Introduction

User interfaces controlled by a virtual agent have begun to be widely researched in recent years. When the user, rather than a computer/algorithm, has the opportunity to control a virtual agent, it is called an avatar.

Our research presents an emotionally-responsive 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 interactions, many emotionally responsive agents only have the ability to have a conversation with the user (they are called embodied conversational agents), e.g., they communicate with users in terms of dialogues and react with behaviors during the conversation. However, they cannot recognize natural gestures from naive users and collaborate with the user to finish a task. In this study, we extracted social cues that did not involve large gestures or solely verbal communication from human pairs that worked 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 emotionally responsive avatar was to work with the person to build structures (a staircase, etc.) out of virtual blocks in a blocks world. The choice of blocks world as the basis for the peer-to-peer human–computer communication highlighted here traces back to the earliest work in Artificial Intelligence (AI) at MIT in the 1960s. In the history of AI blocks world became the focus of some of the earliest work on Computer Vision and planning [1]. 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 use facial expressions linked to certain emotions to motivate the user to express positive emotions. We labelled this version of Diana as Demo. 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 is defined as 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. It was necessary to use a closed environment. Having the human subjects in different rooms for the human-to-human study provided the "display" feedback that Diana provides. One of the participants 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 occurring gestures and used them on training the gesture set that 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 emotional responsiveness. 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 [3]. One of the motivations for the research is the ability to give machines emotional intelligence, including simulating empathy [3]. 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 affect users' interaction experience [4]. This finding reveals the importance of designing human-centered embodied agents that can improve user perception and experience. Concerning the perceived load of humans in the 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, like humans' companions would when the user feels negative emotions.

This research aims to provide an empirical human subject study of Diana with the ability to simulate emotional responsiveness. To compensate for the shortage of studies on emotionally-responsive agents in the computer science field [5], we used psychological theory background as the ground-truth guidance and conducted an interdisciplinary investigation. Diana's gestures were modeled on human naturally occurring gestures, and her facial expressions were synthesized based upon human affects and psychological concepts from Yacin's hierarchical model of empathy [5]. Our study added another step to the research of natural 3D user interfaces in providing a human-centered experience through the use of an emotional avatar.

1.1. Hypothesis

The three states of Diana (Demo, Emotionless, Mimicry) will receive different Likert scale scores in the subjective user experience survey. The three states of Diana cost users different amounts of time to finish the same task. The three states of Diana will receive different scores in the NASA TLX in measuring the task load.

1.2. Facial Expressions Generation

In our work, affect essentially means Diana's ability to recognize user facial expressions as emotions and generate facial expressions meant to reflect responsive emotion.

In Diana's affect module, the actual user emotion detection was implemented by an expression recognition toolkit called Affdex originated from the Affectiva Team [6]. 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 [7] and combined with our addition of some action units.

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 [5]. 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 [8]. 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 [7] to synthesize them into different expressions of affective states.

1.3. 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. 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 [9], an HP 4310 webcam, a Yeti USB microphone, a keyboard, and a mouse. Figure 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 bounded by blue tapes on the ground in front of the table, and it divided out the area (1.6∼2 m, nearest to farthest, −0.8∼0.8 m, 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.

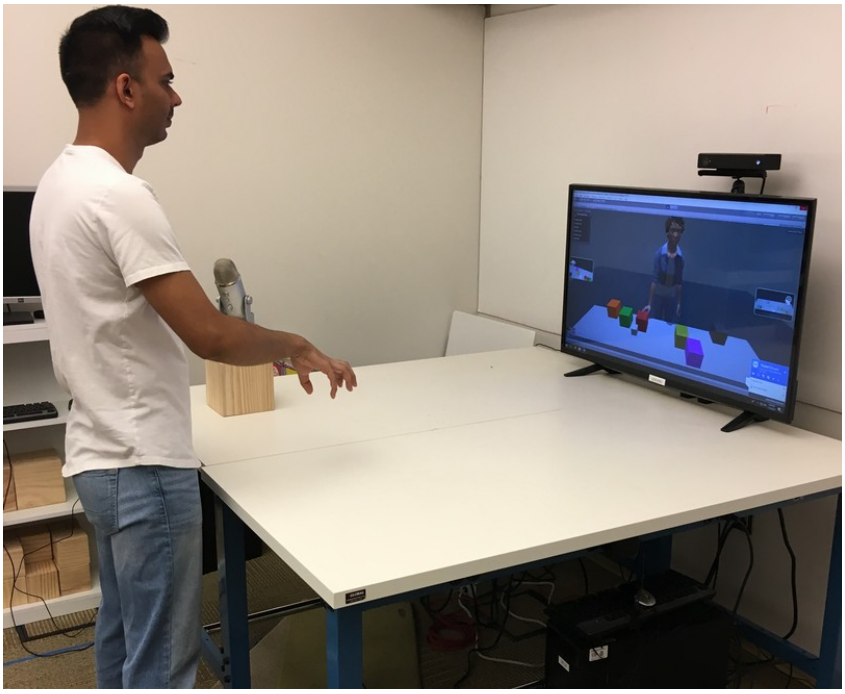
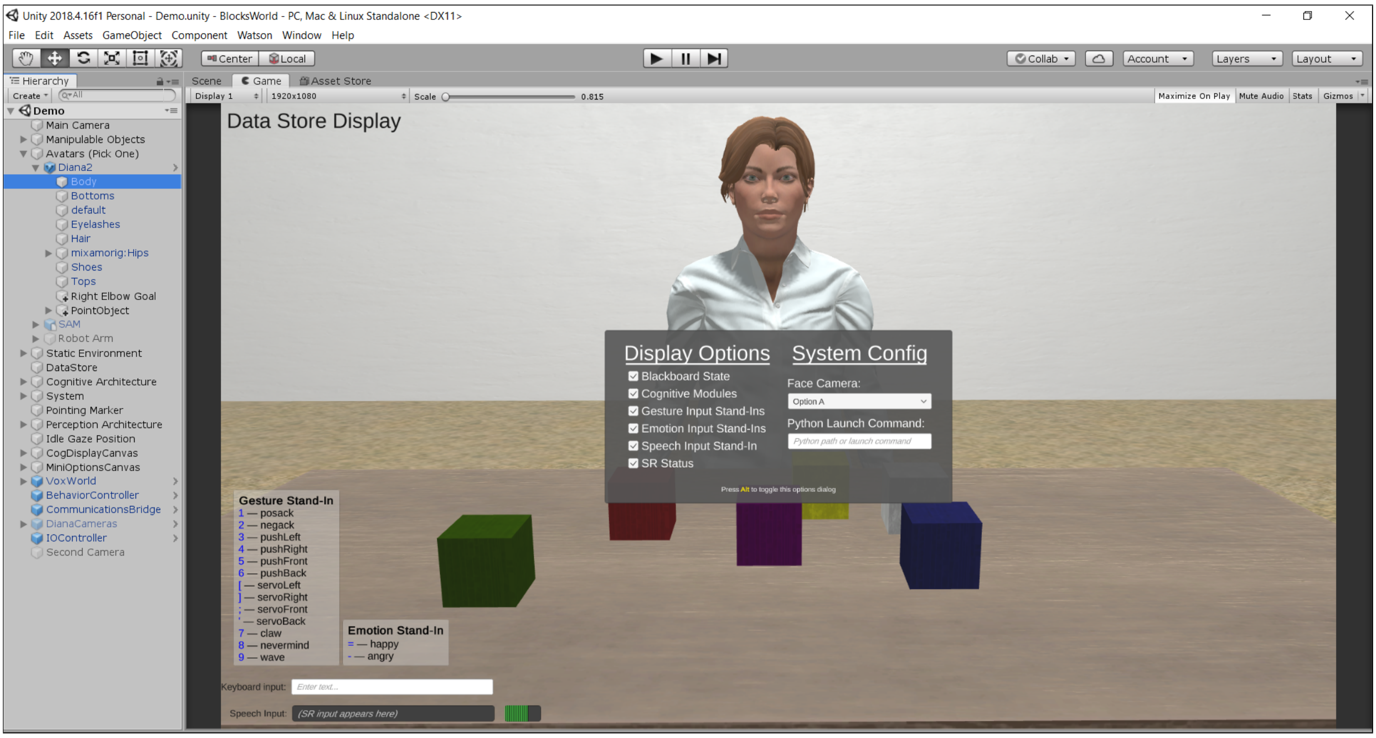


Figure 1. Lab setup used the Kinect v2 sensor for training (Reprinted with permission from J.R. Beveridge, et al. (2019). 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) [10]).

Figure 2 presented the internal framework and interface of the 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 by the Kinect sensor [9] 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 that the user is looking.

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 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 [9] 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.



Data Store Display

Display Options

☑ Blackboard State
☑ Cognitive Modules
☑ Gesture Input Stand-Ins
☑ Emotion Input Stand-Ins
☑ Speech Input Stand-In
☑ SR Status

System Config

Face Camera: Option A

Python Launch Command: Python path or launch command

Press Alt to toggle this options dialog

Gesture Stand-In

1	posack
2	negack
3	pushLeft
4	pushRight
5	pushFront
6	pushBack
[	servoLeft
]	servoRight
;	servoFront
'	servoBack
7	claw
8	nevermind
9	wave

Emotion Stand-In

=	happy
-	angry

Keyboard input: Enter text...

Speech Input: (SR input appears here)

Figure 2. The framework and interface of the Diana system.

Speech recognition was implemented by using the Google Speech-to-Text Engine [11], 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 "CognitiveDisplay-Canvas" 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 3 shows 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. We needed a clear signal to Diana to reverse her actions at the cost of losing some natural interaction qualities.

One characteristic 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 responding to the user's actions. The system efficiency and user engagement could be improved.

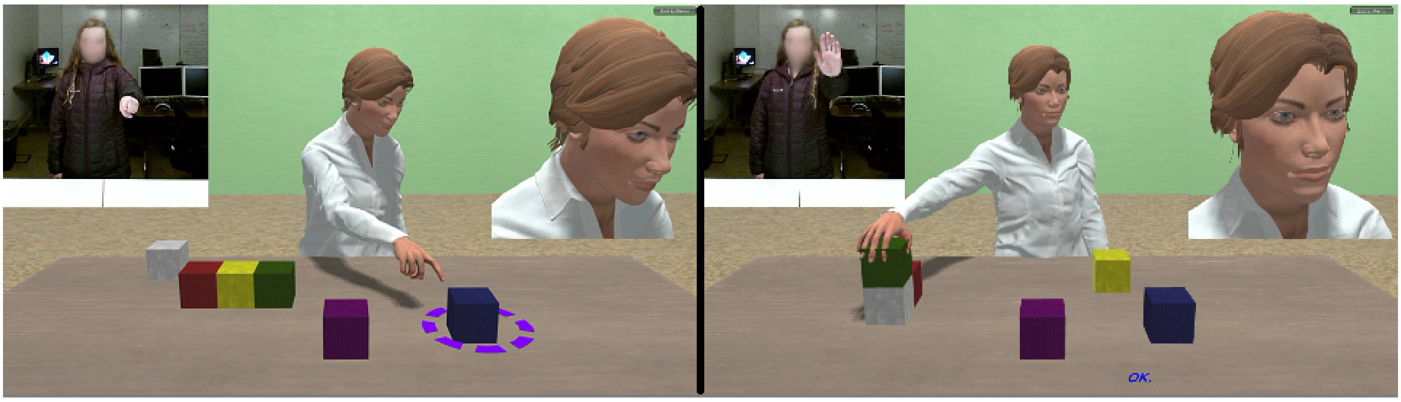


Figure 3. Pointing and never-mind, the two most frequently used gestures during user interaction.

1.4. 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 the EGGNOG dataset [2]. The dataset included over 7 h 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 [12]. For estimating hand pose, we wrote Python clients to take the frame information (such as the depth data provided by the Kinect sensor [9]) 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 [13] to process RGB frames without the depth data.

1.5. Affect Recognition Using the Affdex SDK

Affectiva [6] is a human perception AI company that originated from the MIT Media Lab. It provided a multi-platform SDK named Affdex [13] 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 human behavioral 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 seven 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 accessed on 4 February 2020.

2. Literature Review

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

2.1. Affect in Embodied Agents

Previous studies have found the value of including non-verbal communication channels in embodied agents [14], 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 [15–18]. Compared with an emotionless agents, users rated higher subjective scores when they worked with affective agents [19]. They also spent more time interacting with these agents and indicated they were more willing to use the agent in future interactions [20].

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 [21] 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 [22–24] annotated the agent-client dialogues, and validated the presence of emotions via perceptual tests. Researchers 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 [25] used a technique called Best-Worst Scaling [26] to create an emotion intensity dataset. 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 [27] presented a protocol for the collection and annotation of multimodal emotional behaviors (speech and gestures) that 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 [14] 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. 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. [28] 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. [29], 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.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. [19] 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 emotionally responsive avatar that randomly nodded when the participant was speaking and mimicked the participant's facial expressions. Results indicated the emotionally responsive avatar with the facial expression model received an increase in the positivity of users' ratings.

Similarly, Aneja, McDuff, and Shah [30] 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. 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. [31] 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. 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. Emotional Responsiveness in Embodied Agents

Our avatar was modeled from human behaviors aimed at simulating a more complex affect to improve the user's perception in tasks. Though affect has been intensively studied in Affective Computing, it is still challenging to create an emotionally intelligent embodied agent. Empathy has been defined in the scientific literature as the capacity to relate to another's emotional state and has been assigned to a broad spectrum of cognitive and behavioral abilities [5]. Agents merely mimicking human facial expressions are in the fundamental level of a hierarchical empathy model and only represent a low-level empathic behavior towards users [5], 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. [20] designed a learning software with an upper-body virtual agent on the side of the window to promote students' engagement and enhance learning. 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 [32] 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. 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. [8] at the University of Glasgow found the standardized facial expressions previously thought to be 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 4. 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 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.

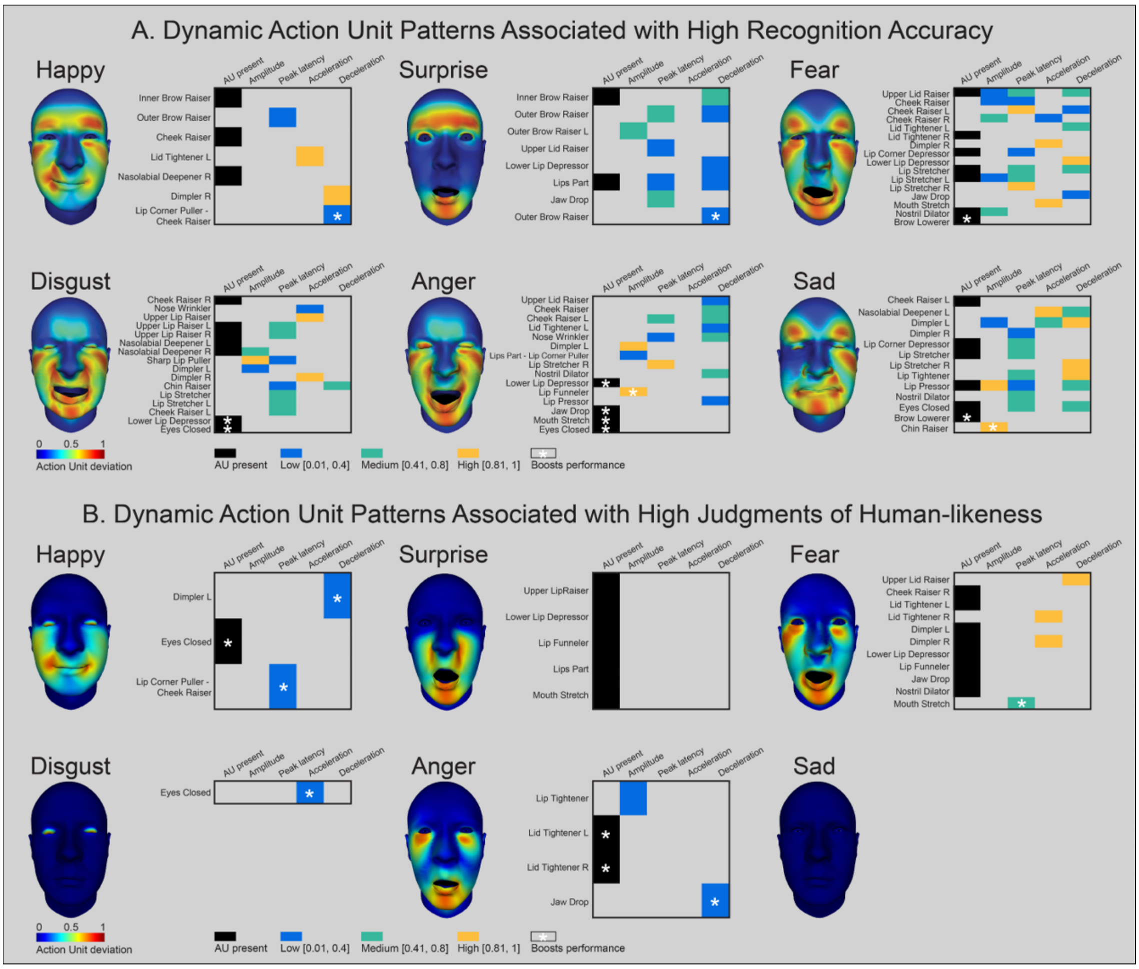


Figure 4. The culturally-sensitive dynamic action units that are associated with high recognition accuracy in panel (**A**) and high judgments of human-likeness in panel (**B**) (Reprinted with permission from C. Chen, et al. (2019). 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition Name) [8].)

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 [5] and created three modes of Diana to give a comprehensive comparison of their ability to improve user experience, as shown in Figure 5. The three modes included an Emotionless avatar that was commonly studied in commercial activities and research, the Mimicry avatar with the 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 to simulate high-level emotional responsiveness.

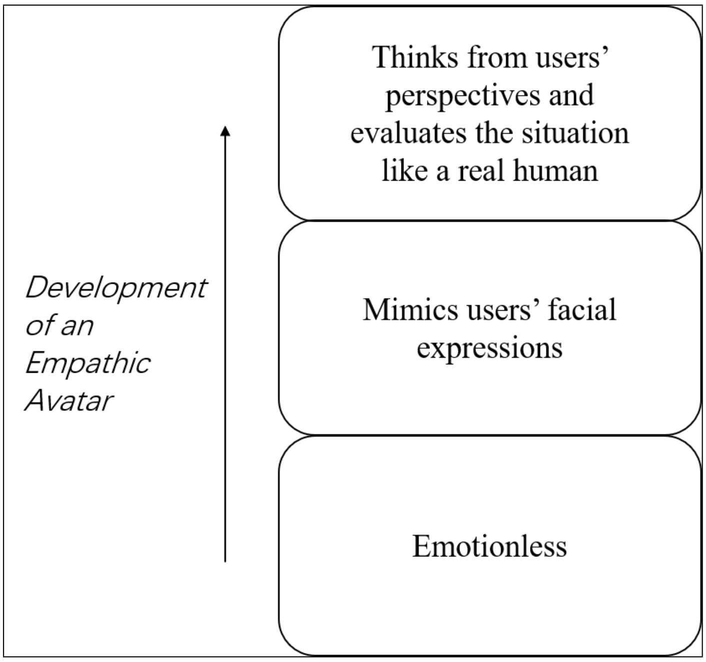

Development of an Empathic Avatar
Thinks from users' perspectives and evaluates the situation like a real human
Mimics users' facial expressions
Emotionless

Figure 5. Three modes of Diana following a hierarchical model of empathy.

We 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 as 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 emotionally intelligent agent. Cohn [33] 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 the interaction. In the study by Chen et al. [20], while most of the subjects expressed their willingness to continue the interaction, some subjects also reported the empathic tutor had elicited more frustration and worry during the learning process. Another study by Haring et al. [34] 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 [35]. To generate humanoids expressions, many researchers have provided diverse solutions. Belhaj, Kebair, and Said [36] 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. [37] conducted a study to investigate some action units or their combinations that were most likely to be recognized as certain emotions under human appraisals. 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 [38]. 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 [39] also illustrated the major challenges in building computational models of emotions for autonomous agents and presented a novel approach. The challenges included the integration of cognition and emotions in agent architectures, the 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 three-layer integrative framework and a general methodology to guide the development of biologically inspired Computational Models of Emotions within three 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.

3. Materials and Methods

This section 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 emotionally responsive behaviors.

3.1. 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 four 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 6. For example, "Which most expresses JOY?" paired with Figure 6 was one of the questions asked of participants. Figures 7–10 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 who chose each level were nearly equally-distributed. Considering the analysis of our result, we set up Diana's joyful and sympathetic facial expressions to have nearly 100% of the 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.

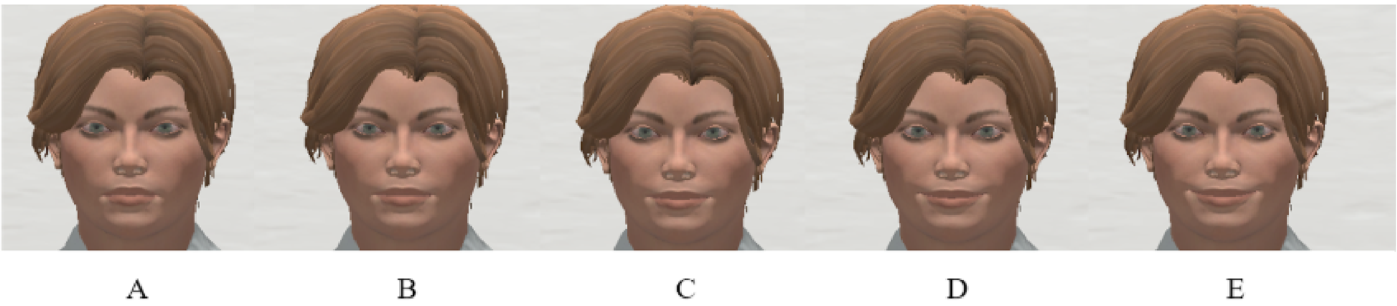


Figure 6. Diana's joy from level (**A**–**E**) intensified from 20% to 100% of the overall scale.

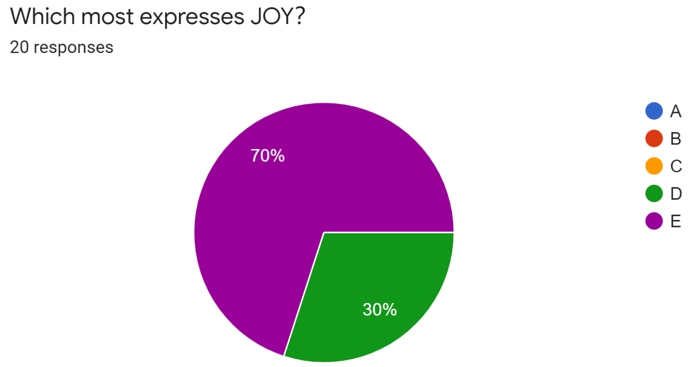


Figure 7. Joy.

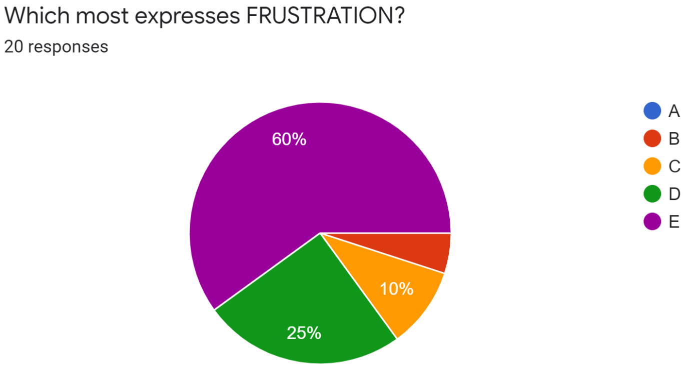


Figure 8. Frustration.

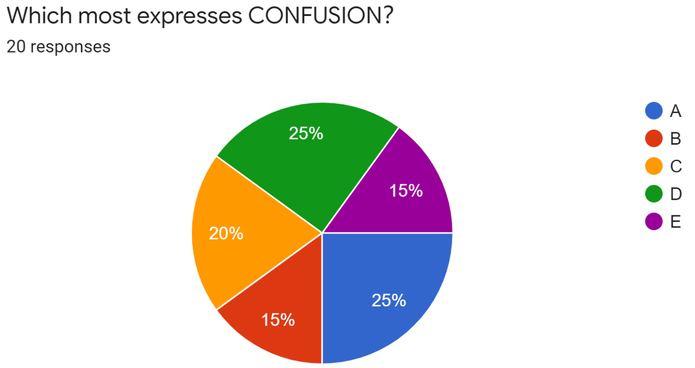


Figure 9. Confusion.

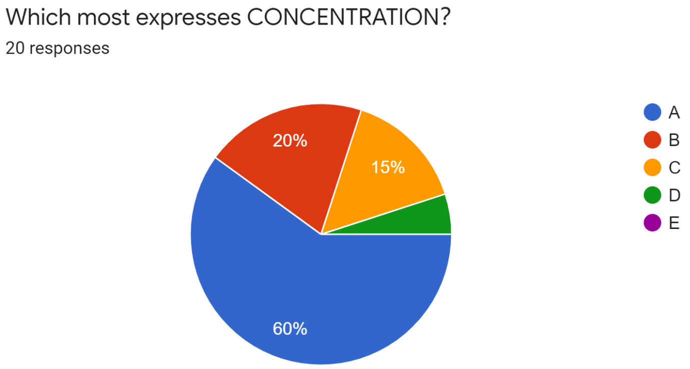


Figure 10. Concentration.

3.2. 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 emotionally intelligent 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 is composed of a finite state machine [40] 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 s.

Specifically, in the Demo mode of Diana, different than the agent in the iViz Lab who expressed emotional intelligence 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 [40]. 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 11. The orange lines represented conditions relevant to the user's affects and the blue lines represented 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.

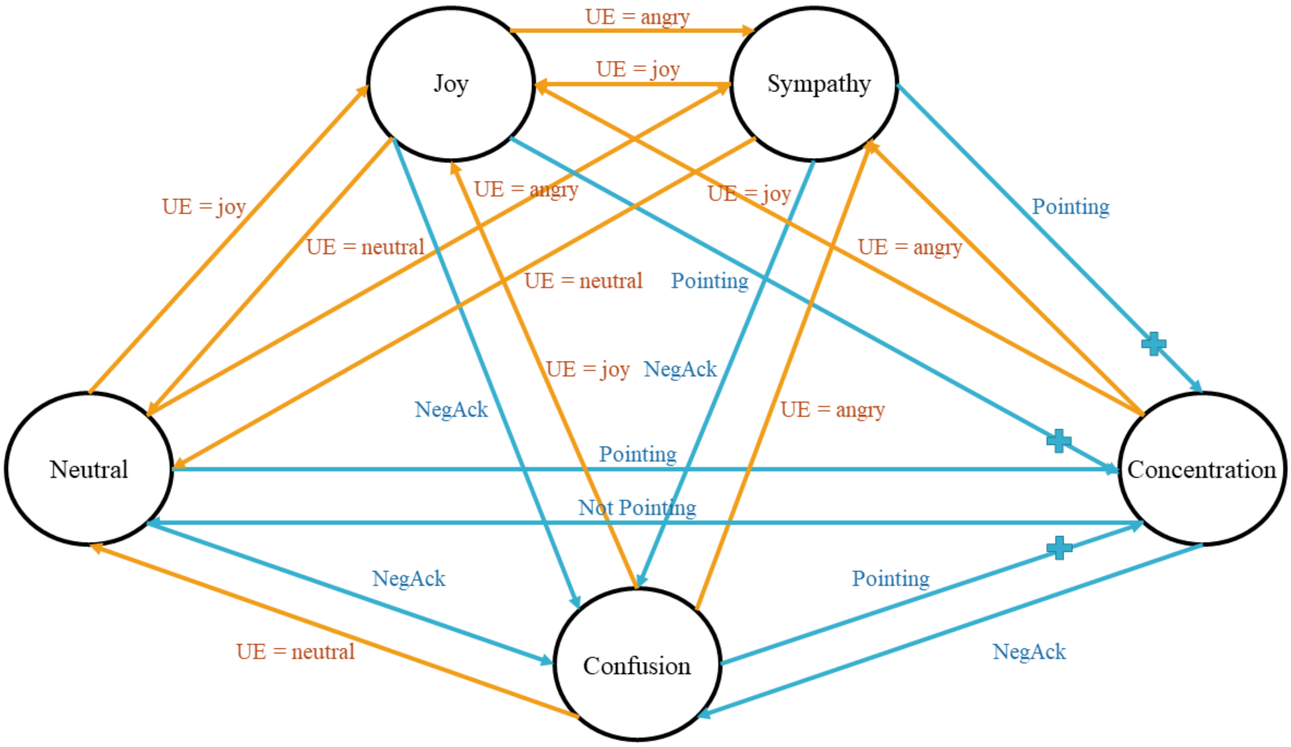


Figure 11. A diagram of the finite state machine of Diana's affective states.

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 Algorithm 1 which updates Diana's responsive affect transitions between five states is shown below.

Algorithm 1: Algorithm updating Diana's responsive affect transitions.

```
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
```

3.3. 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 [7] seven basic emotions, see [8,19,20,28,29,32]. 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 [5], 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. [5].

Table 1 showed Diana's action code combinations for expressions compared with the code combinations in a standard Facial Action Coding System [7] and SmartBody [41]. 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 [7]. 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 [8]. 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 four non-neutral affective states are shown in Figures 12–15.

Table 1. Diana's action unit code combinations compared with other code combinations.

Affective States	FACS [7]	SmartBody [41]	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 [42]	-	BrowsIn + Squint + NoseScrunch + JawDown
Concentration	-	-	BrowsUp + EyesWide

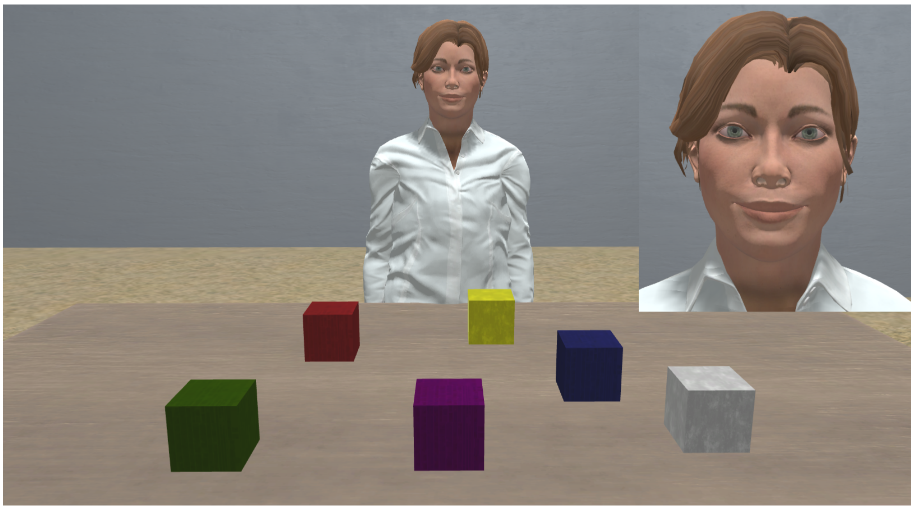


Figure 12. Joy.

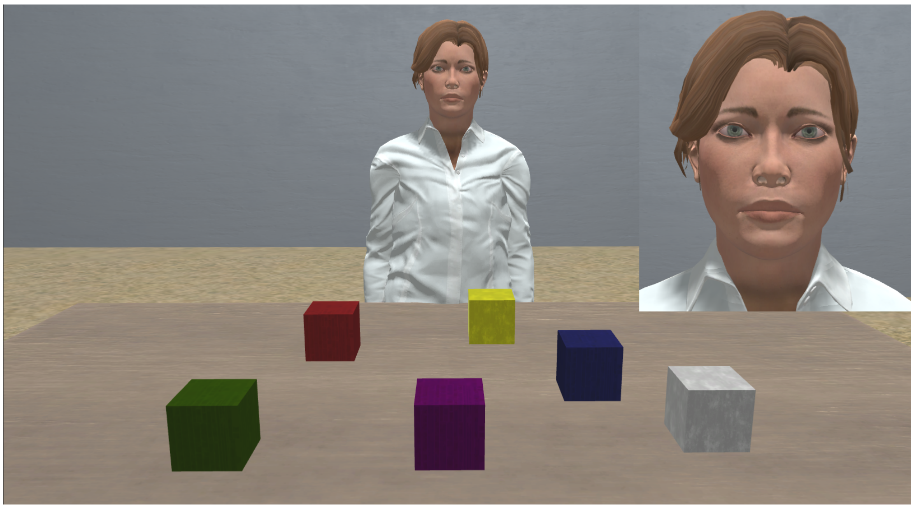


Figure 13. Sympathy.

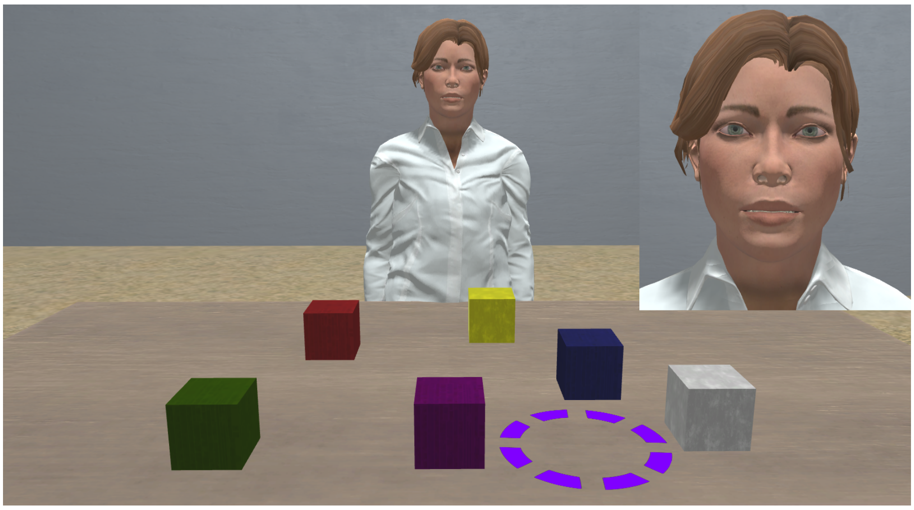


Figure 14. Confusion.

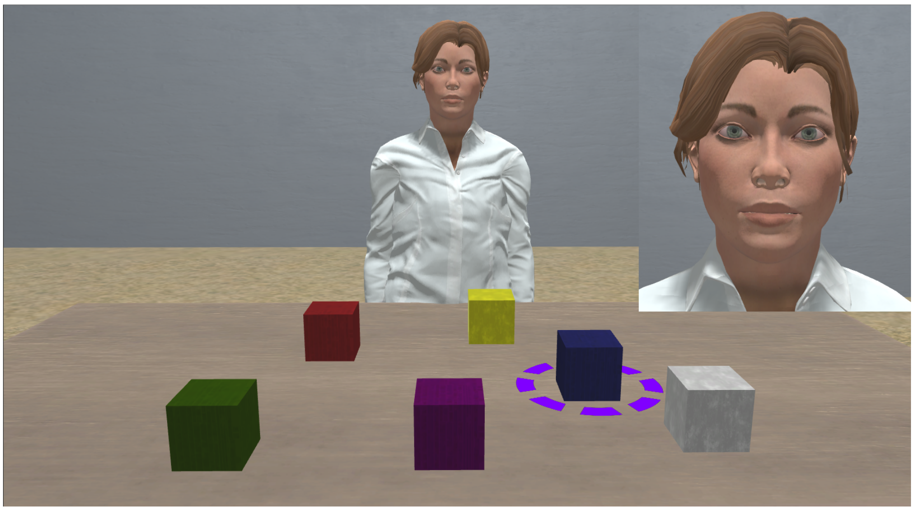


Figure 15. Concentration.

3.4. Experimental Setup

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

3.4.1. Participants

All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with institutional review board (IRB) guidelines, and the protocol was approved by the IRB Committee of Colorado State University (19-9076H) on 10 July 2019.

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 [9]), within these people, six used to play games a lot but in recent years they had shortened the time to 1–2 h a week.

3.4.2. 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 3.6 GHz 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 [9] 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.

3.4.3. 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 (we found that beards interfere with the facial recognition software), and game usage, etc. Then the participants were shown a 3-min 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 the 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 and they were randomized to avoid ordering effects, 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 [43]. 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 five-point Likert scale questionnaire and a NASA Task Load Index survey [44]. 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 five-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 an optional text box provided for subjects to type in comments. In this questionnaire, we asked seven unbiased questions investigating users' perceptions of the whole character's movements and appearance including three 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. Results

4.1. Positive Responses

The raw count of positive responses each mode of Diana received on each question were shown in Table 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.

Table 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 No Affect Mimic User's Affect	D E M	11 10 10
2	It helped to look at the avatar's face	Empathic Affect No Affect Mimic User's Affect	D E M	4 3 3
3	The avatar was helping me	No Affect Mimic User's Affect Empathic Affect	E M D	10 9 7
4	The avatar's facial expressions are natural	Empathic Affect No Affect Mimic User's Affect	D E M	8 7 6
5	The avatar's movements are natural	Empathic Affect Mimic User's Affect No Affect	D M E	10 7 6
6	I felt comfortable working with this avatar	Empathic Affect No Affect Mimic User's Affect	D E M	13 12 11
7	I felt relaxed working with this avatar	Empathic Affect Mimic User's Affect No Affect	D M E	10 10 9

4.2. Distribution of Ratings Votes

Figure 16 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 the 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.

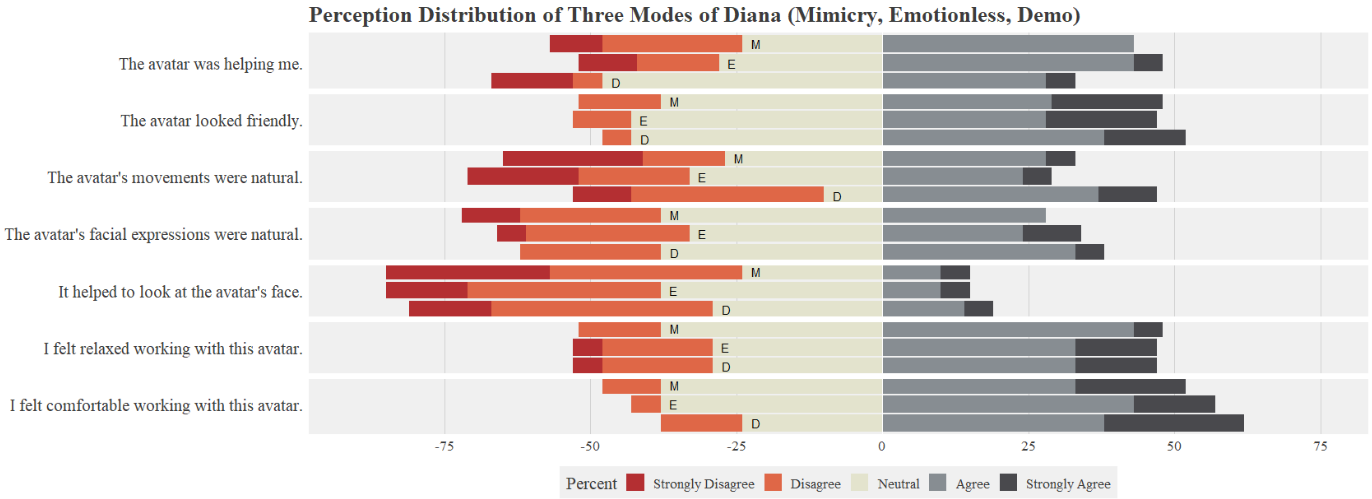


Figure 16. Diverging stacked bar chart of the percentages of users' perceptive votes.

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 showed no significant difference between the votes in three modes of Diana, as the statistics and p values shown in Table 3.

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. However, 3 participants also mentioned the Uncanny Valley effect [35] and reported the avatar's facial expressions made them think they did something wrong.

Table 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

4.3. 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 [44]. In our experiment, we used it as a survey to measure six metrics (mental demand, physical demand, temporal demand, performance, effort, and frustration) of the task. Table 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 5. It meant the perceived workload of three modes of Diana were the same.

Table 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 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

4.4. 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 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 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

5. Discussion

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 the 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 dependent 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. We asked participants 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 emotionally responsive facial expressions.

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, which 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 [7]. 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 was 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. This may include a set of toys on top of a table while keeping the same set of toys in the virtual world. A human will move around the tasks to explain to the embodied agent how to move them. A follow-up may include the use of augmented reality headsets to improve engagement.

6. Conclusions 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 avatars 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. 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 utilized our previous 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 emotional intelligence 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 [7], 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 who 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 was 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 five-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 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 emotionally intelligent 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 avatar-human 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 using avatars.

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 research 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 emotionally responsive avatar. Both positive and negative votes indicated our approach of adding emotionally intelligent 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, letting 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 emotionally intelligent 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 human affect.

Author Contributions: Conceptualization, H.W., J.R.B., and F.R.O.; methodology, H.W. and J.R.B.; software, H.W.; validation, H.W., F.R.O., J.R.B. and V.G.; formal analysis, H.W.; investigation, H.W.; resources, J.R.B.; data curation, H.W., J.R.B. and V.G.; writing—original draft preparation, H.W.; writing—review and editing, V.G.; visualization, H.W.; supervision, F.R.O. and J.R.B.; project administration, J.R.B., F.R.O. and V.G.; funding acquisition, J.R.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the DARPA Communicating with Computers program under ARO contract W911NF-15-1-0459.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Colorado State University (protocol code: 19-9076H and date of approval: 7 October 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data discussed in this paper can be found at: http://nrdc.cs.colostate.edu/nuidata (accessed on 4 March 2020).

Acknowledgments: We would like to express a deep appreciation for the support of Julia Sharp and Bruce Draper. We would like to thank the vision lab members 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.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Slaney, J.; Thiébaux, S. Blocks world revisited. *Artif. Intell.* **2001**, *125*, 119–153. [CrossRef]
2. Wang, I.; Fraj, M.B.; Narayana, P.; Patil, D.; Mulay, G.; Bangar, R.; Beveridge, J.R.; Draper, B.A.; Ruiz, J. EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, USA, 30 May–3 June 2017; pp. 414–421.
3. Heylen, D.; Ghijsen, M.; Nijholt, A.; op den Akker, R. Facial signs of affect during tutoring sessions. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Beijing, China, 22–24 October 2005; Springer: Berlin/Heidelberg, Germany, 2005; pp. 24–31.
4. Rickenberg, R.; Reeves, B. 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, The Hague, The Netherlands, 1–6 April 2000; pp. 49–56.
5. Yalcin, Ö.N.; DiPaola, S. A computational model of empathy for interactive agents. *Biol. Inspired Cogn. Archit.* **2018**, *26*, 20–25. [CrossRef]
6. Affectiva Human Perception AI Analyzes Complex Human States. Available online: https://www.affectiva.com (accessed on 4 February 2021).
7. Ekman, P.; Friesen, W.V.; Hager, J.C.; Firm, A.H.F. *Facial Action Coding System*; A Human Face: Salt Lake City, UT, USA, 2002.
8. Chen, C.; Hensel, L.B.; Duan, Y.; Ince, R.A.; Garrod, O.G.; Beskow, J.; Jack, R.E.; Schyns, P.G. Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–8.
9. Kinect for Windows. Available online: https://developer.microsoft.com/en-us/windows/kinect (accessed on 4 March 2020).
10. Mulay, G.; Draper, B.A.; Beveridge, J.R. Adapting RGB Pose Estimation to New Domains. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 7–9 January 2019; pp. 324–330.
11. Speech-to-Text Accurately Convert Speech into Text Using an API Powered by Google's AI Technologies. Available online: https://cloud.google.com/speech-to-text (accessed on 4 February 2021).
12. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
13. McDuff, D.; Mahmoud, A.; Mavadati, M.; Amr, M.; Turcot, J.; Kaliouby, R.E. AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, San Jose, CA, USA, 7–12 May 2016; pp. 3723–3726.
14. Marsi, E.; Van Rooden, F. Expressing uncertainty with a talking head in a multimodal question-answering system. In Proceedings of the MOG 2007 Workshop on Multimodal Output Generation, Aberdeen, UK, 25–26 January 2007; p. 105.
15. Van Mulken, S.; André, E.; Müller, J. The persona effect: How substantial is it? In *People and Computers XIII*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 53–66.
16. Baylor, A.L.; Kim, Y. Simulating instructional roles through pedagogical agents. *Int. J. Artif. Intell. Educ.* **2005**, *15*, 95–115.
17. Kipp, M.; Kipp, K.H.; Ndiaye, A.; Gebhard, P. Evaluating the tangible interface and virtual characters in the interactive COHIBIT exhibit. In Proceedings of the International Workshop on Intelligent Virtual Agents, Marina del Rey, CA, USA, 21–23 August 2006; Springer: Berlin/Heidelberg, Germany, 2006; pp. 434–444.
18. Fan, L.; Scheutz, M.; Lohani, M.; McCoy, M.; Stokes, C. Do we need emotionally intelligent artificial agents? First results of human perceptions of emotional intelligence in humans compared to robots. In Proceedings of the International Conference on Intelligent Virtual Agents, Stockholm, Sweden, 27–30 August 2017; Springer: Cham, Switzerland, 2017; pp. 129–141.
19. Pablos, S.M.; García-Bermejo, J.G.; Zalama Casanova, E.; López, J. Dynamic facial emotion recognition oriented to HCI applications. *Interact. Comput.* **2013**, *27*, 99–119. [CrossRef]
20. Chen, G.D.; Lee, J.H.; Wang, C.Y.; Chao, P.Y.; Li, L.Y.; Lee, T.Y. An empathic avatar in a computer-aided learning program to encourage and persuade learners. *J. Educ. Technol. Soc.* **2012**, *15*, 62–72.
21. Pantic, M.; Pentland, A.; Nijholt, A.; Huang, T.S. Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 47–71.
22. Devillers, L.; Vasilescu, I.; Lamel, L. Annotation and detection of emotion in a task-oriented human-human dialog corpus. In Proceedings of the ISLE Workshop, Bendor, France, 10–12 July 2002.
23. Devillers, L.; Lamel, L.; Vasilescu, I. Emotion detection in task-oriented spoken dialogues. In Proceedings of the 2003 International Conference on Multimedia and Expo. ICME '03, (Cat. No.03TH8698), Baltimore, MD, USA, 6–9 July 2003; Volume 3, p. III-549.

24. Devillers, L.; Vasilescu, I.; Mathon, C. Prosodic cues for perceptual emotion detection in task-oriented Human-Human corpus. In Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 3–9 August 2003.
25. Mohammad, S.M.; Bravo-Marquez, F. WASSA-2017 shared task on emotion intensity. *arXiv* **2017**, arXiv:1708.03700.
26. Louviere, J.J.; Flynn, T.N.; Marley, A.A.J. *Best-Worst Scaling: Theory, Methods and Applications*; Cambridge University Press: Cambridge, UK, 2015.
27. Zara, A.; Maffiolo, V.; Martin, J.C.; Devillers, L. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Lisbon, Portugal, 12–14 September 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 464–475.
28. Ku, J.; Jang, H.J.; Kim, K.U.; Kim, J.H.; Park, S.H.; Lee, J.H.; Kim, J.J.; Kim, I.Y.; Kim, S.I. Experimental results of affective valence and arousal to avatar's facial expressions. *CyberPsychol. Behav.* **2005**, *8*, 493–503. [CrossRef]
29. Shen, J.; Rudovic, O.; Cheng, S.; Pantic, M. Sentiment apprehension in human-robot interaction with NAO. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 867–872.
30. Aneja, D.; McDuff, D.; Shah, S. A High-Fidelity Open Embodied Avatar with Lip Syncing and Expression Capabilities. In Proceedings of the 2019 International Conference on Multimodal Interaction, Suzhou, China, 14–18 October 2019; pp. 69–73.
31. Mudrick, N.V.; Taub, M.; Azevedo, R.; Rowe, J.; Lester, J. Toward affect-sensitive virtual human tutors: The influence of facial expressions on learning and emotion. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 184–189.
32. Esteban, P.G.; Insua, D.R. An Affective Model for a non-Expensive Utility-based Decision Agent. *IEEE Trans. Affect. Comput.* **2019**, *10*, 498–509. [CrossRef]
33. Cohn, J.F. Foundations of human computing: Facial expression and emotion. In Proceedings of the 8th International Conference on Multimodal Interfaces, Banff, AB, Canada, 2–4 November 2006; pp. 233–238.
34. Haring, K.; Nye, K.; Darby, R.; Phillips, E.; de Visser, E.; Tossell, C. I'm Not Playing Anymore! A Study Comparing Perceptions of Robot and Human Cheating Behavior. In Proceedings of the International Conference on Social Robotics, Madrid, Spain, 26–29 November 2019; Springer: Cham, Switzerland, 2019; pp. 410–419.
35. MacDorman, K.F.; Ishiguro, H. The uncanny advantage of using androids in cognitive and social science research. *Interact. Stud.* **2006**, *7*, 297–337. [CrossRef]
36. Belhaj, M.; Kebair, F.; Said, L.B. Emotional dynamics and coping mechanisms to generate human-like agent behaviors. *Appl. Artif. Intell.* **2017**, *31*, 472–492. [CrossRef]
37. Scherer, K.R.; Mortillaro, M.; Rotondi, I.; Sergi, I.; Trznadel, S. Appraisal-driven facial actions as building blocks for emotion inference. *J. Personal. Soc. Psychol.* **2018**, *114*, 358. [CrossRef] [PubMed]
38. Scherer, K.R.; Schorr, A.; Johnstone, T. *Appraisal Processes in Emotion: Theory, Methods, Research*; Oxford University Press: Oxford, UK, 2001.
39. Rodríguez, L.F.; Ramos, F. Computational models of emotions for autonomous agents: Major challenges. *Artif. Intell. Rev.* **2015**, *43*, 437–465. [CrossRef]
40. Wang, J. *Formal Methods in Computer Science*; CRC Press: Boca Raton, FL, USA, 2019.
41. SmartBody. Available online: https://smartbody.ict.usc.edu (accessed on 4 February 2021).
42. Tools for the Facial Action Coding System (FACS). Available online: https://www.noldus.com/applications/facial-action-coding-system (accessed on 4 February 2021).
43. Permuted-Block-Randomization. Available online: https://www.statisticshowto.com/permuted-block-randomization (accessed on 4 February 2021).
44. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.