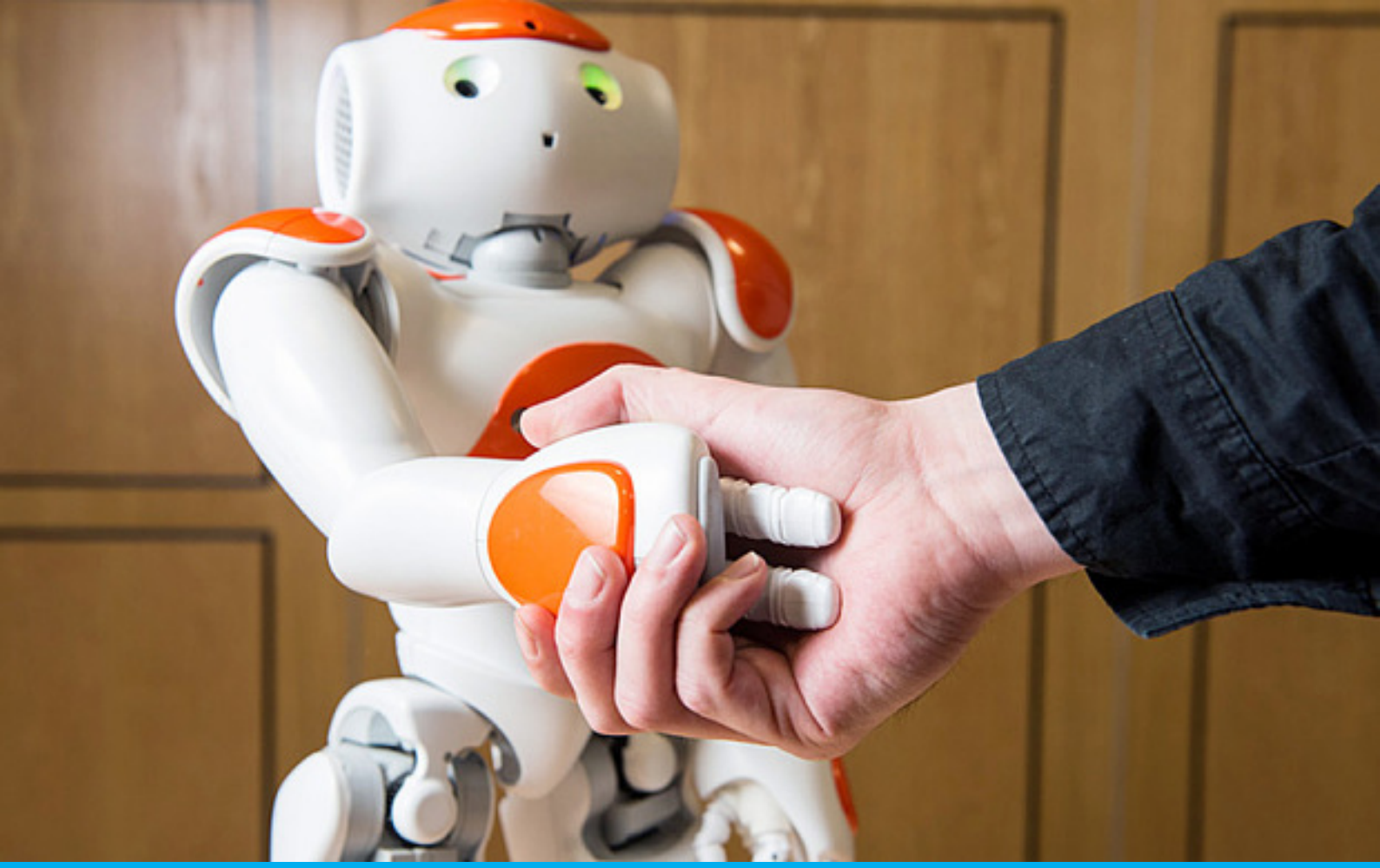


Affect expression in social robots

Combining non-verbal affect expression techniques

Pooja Prajod

Technische Universiteit Delft

TU Delft

Affect expression in social robots

Combining non-verbal affect expression techniques

by

Pooja Prajod

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
to be defended on Wednesday August 28, 2019 at 02:00 PM.

Student number: 4725522
Project duration: December 5, 2018 – July 17, 2019
Thesis committee: Dr. Koen V. Hindriks, TU Delft, supervisor
Dr. Myrthe L. Tielman, TU Delft
Dr. Nava Tintarev, TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
Delft University of Technology

TU Delft

Preface

Throughout the Master's programme, I found that human-computer interactions, especially social robotics and affective computing, always piqued my interest. The overlap of psychological concepts and artificial intelligence fascinates me. So I decided to choose a thesis project in the field I knew I would enjoy.

With the advent of social robots which are designed to be 'social', human-like interactions have become a necessity. It is natural for us to use a plethora of emotions to convey additional information or to make an interaction more engaging. But emotion expression is not commonly associated with robots. Many humanoid robots cannot generate facial expressions to portray various emotions. Studies have shown that robots are multi-modal systems which can employ multiple channels to express an emotion. Through this thesis, I explored the non-verbal emotion expression techniques and their expressive capabilities. I found that some emotions are easier to express than others, and a single technique cannot express all the emotions. I chose a few emotions and systematically determined the best technique for each of them.

I would like to thank my supervisor Dr. Koen V. Hindriks, for his guidance and insights throughout this thesis. I would also like to express my gratitude to Ruud de Jong for helping me with the experimental setup on multiple occasions. I am grateful for all the love and support from my family and friends in Delft and back home in India. A special thanks to my boyfriend Sanchar Sharma, who always brought out the best in me.

Pooja Prajod
Delft, July 2019

Contents

1

Introduction

From greeting customers to health care assistants, social robots are becoming an integral part of our lives. The word 'social' implies that, in addition to performing various tasks, these robots are designed to integrate with our society. Interacting with humans is an essential part of being social. [8] recognises two channels in human interactions: explicit or the actual message and, implicit or the information about the speaker. Affect (mood, emotion, etc.) is a vital component of the implicit channel. Thus, holistic human-computer interactions would include recognising the expressed affects, as well as expressing own affect. Though researches in both aspects employ similar concepts, the challenges involved are different. This thesis focuses on equipping social robots with the capability of expressing affects. Some of the channels that can be explored for expressing affects are:

1. Facial expressions like emoticons
2. Actual dialogue spoken by the robot
3. Voice features like volume, pitch, etc.
4. Body language and pose
5. Motion features like speed, acceleration, etc.
6. Coloured pattern display like LEDs

This thesis explores channels pertaining to motion, body-language, LED colour and pose.

1.1. Affect representation

In [36], the authors highlight seven emotions: neutral, happy, sad, anger, disgust, fear, surprise, and four cognitive states: interested, bored, frustrated and puzzled. Indian classical dances with an emphasis on story-telling use 9 principle emotions or Navarasas including surprise, happy, sad, anger, peace, love, disgust, courage and fear. Though there are overlaps between the affects identified in different domains, it can be observed that the categories of affects are finite but not fixed. Many studies including [5, 11, 26] represents affect using three dimensions: *valence, arousal, dominance*. Valence can be positive or negative, representing levels of pleasure ranging from *unpleasant* to *pleasant*. Arousal signifies the energy of the affect ranging from *un-aroused or calm* to *aroused or exited*. Dominance indicates the level of control and ranges from *no control* to *full control*. This thesis uses a simplified 2D representation of affect involving only valence and arousal.

1.2. Challenges and Constraints

Facial expressions play an important role in expressing affect. While some robots such as Robotinho [20] or iCat [7] try to mimic human expressions using facial features, many robots

like NAO lack such capabilities. In such cases, a popular approach is to mimic distinct key poses and gestures that are often associated with specific emotions. For example, an expressive medium like emoticon can easily portray sadness through tears and inverted smile, whereas a NAO robot resorts to a head-down pose. A limitation of such an approach is that it often interrupts the task being performed. Consider an example of nodding 'yes' by moving the head up and down. Expressing sad by a head-down pose would interfere with the nodding task since both use the same joints.

With ever-increasing types and models of social robots, it is not efficient to have dedicated studies and systems for each of them. Studies like [5, 16, 30] demonstrated that some motion and body language features could be modified to express the mood or state of a robot. These solutions are independent of the robot or the task being performed. This was the motivation for developing a generic affect expression framework that modifies robot-independent features without altering or interrupting the tasks.

Robots are often designed for specific domains or tasks. For example, robot arms are mainly designed for industrial purposes and have a very limited scope of affect expression. While some of the features explored in this thesis may apply to various robots, the focus of the study is limited to humanoid robots with minimal body features, which can perform human-like motions.

1.3. Problem statement

The goal of this thesis is to build a *generic parametric* framework to express affect. The framework is generic because it would use robot-independent non-verbal features, and thus can be used for expressing affects in many simple humanoid robots. The framework is parametric because it would generate affective gestures for any point on the valence-arousal plane.

1.3.1. Affect list

This work studies the expression of affects in the valence-arousal plane. But it is impossible to test all the feature modulations in a continuous space like the valence-arousal space. Hence, the experiments are limited to a concise list of discrete points covering different parts of the valence-arousal space.

Plutchik [23] proposed a multi-dimensional emotion representation which identifies eight basic emotions: joy, sadness, anger, fear, trust, disgust, surprise and anticipation. Other emotions are viewed either as an intensity-variant of these emotions or as a combination of two or more basic emotions. In [24], Russell plots various affects on the valence-arousal plane, forming approximately a circle. This work focuses on expressing the affects: *happy, excited, anger, fear, sad, tired, relaxed* and *content*. These affects fall into different categories in Plutchik's wheel of emotions and can be roughly mapped on the valence-arousal plane as seen in figure 1.1. These affects cover all four quadrants formed by the valence and arousal axes. There are multiple pairs of affects in this set which have similar valence but different arousal values or vice versa. For example, positive valence pairs like excited-relaxed or negative valence pairs like sad-fear have the same valence values but differ in arousal. Similarly, pairs like anger-excited or tired-relaxed have same arousal levels but different valences. Additionally, each quadrant has affects like anger-fear, which may be difficult for the users to distinguish because they have the same signs for valence and arousal.

1.3.2. Evaluation

The research questions of this thesis pertain to the perceived affect of a gesture performed by the robot. We use two methods to collect the data about the perceived affect: emotion labels and valence-arousal ratings. The perceived valence and arousal of the expressed affect are rated by the participants using a 7-point SAM(Self-Assessment Manikin) [4] questionnaire. In addition, the participants have to choose an emotion from the given list (Neutral, Excited, Happy, Content/Satisfied, Relaxed, Tired, Sad, Fear and Angry) which they think best rep-

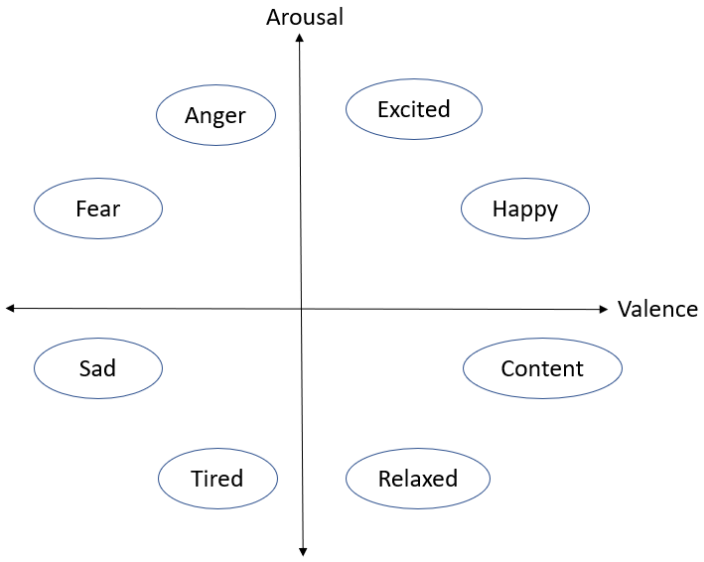


Figure 1.1: A simplified adaptation of Russell's circular model showing the mapping of affects focused in this thesis.

resents the affect expressed by the robot.

These two kinds of data are analysed to draw conclusions about the model and to judge whether the expressed affects are distinguishable and perceived as intended. We use emotion labels to calculate the recognition rate of the affect and classify it as *low, medium* or *high*. Let's say, each participant has a probability p of recognising the expressed affect. Since the observations are independent and identical, they can be seen as Bernoulli trials. Using maximum likelihood estimation of p, it can be derived that $p = \frac{r}{N}$, where r is the number of participants who recognised the expressed affect and N is the total number of participants. By setting p threshold to 0.75 for high and 0.5 for medium, the recognition rate can be classified as seen in table 1.1.

Percentage of participants	Identification rate
≥ 75%	High
50 - 75 %	Medium
< 50 %	Low

Table 1.1: Recognition rate classification based on the chosen thresholds

The perceived valence-arousal data gives more insight into the affect expression capability of the models. The affect expression models are built on a valence-arousal based model. So it is crucial to investigate the valence-arousal ratings to verify if the perceived valence and arousal are close to the intended values. Affects close to the intended values are promising candidates for further analysis. Among the chosen affects, a statistical test is run between neighbouring affects to determine if they are significantly different at least along one of the axes. Affects which are close to the intended values and show significant difference are deemed distinguishable.

This thesis uses an affect-list based evaluation, i.e. the framework is evaluated based on the perception and recognition of the expressed affect. Though this is a commonly used evaluation method, there are a couple of alternatives. In [13], the participants were presented a few fixed head positions. They were asked to choose the head position they would associate with an affect. The affect is considered to be expressible if it is quite often associated with a particular head position. For a framework which modulates several features, this method

may not be suitable. It requires the participants to design the affective gesture using the given set of feature variations. Depending on the number of features and distinct feature values, the possible permutations could be overwhelming. The second method involves collecting valence and arousal ratings for all the variations of a single feature, which then used to map the feature values to emotion labels. This procedure is repeated for all the features. In this method, the effect of combining features is ignored.

1.3.3. Research questions and hypotheses

Given the valence-arousal values of an affect, the main goal of the framework is to express affect by modulating various features without interrupting the task. Some studies like [30] have proposed a parametric model for expressing affect. However, the range of affects expressed depends on the affect model and the variations that can be produced by the underlying modulation functions. For example, [33] focuses on a valence oriented representation and hence primarily expresses happy and sad. As seen from figure 1.1, fear and sad have similar valence values and hence such a model always expresses sadness for negative valence and as a consequence, it is less suitable to express fear. Similarly, [34] explores an arousal oriented approach and found that it could express excited and calm to some extent. Again as seen from figure 1.1, excited and anger have similar arousal values and thus such a model does not differentiate between them. These examples demonstrate that one-dimensional modelling is not sufficient to express the various basic emotions. [30] proposes a model which utilizes both valence and arousal and successfully expresses a few more affects. This model proposed modulating motion and body language features like speed, amplitude, etc. to express mood. The main idea behind such an approach is to modulate the motion and body language features without changing or interrupting the task being performed. This leads us to the first research question:

[Design] 1(a). What are the motion and body language features and the associated operators that can be used in a parametric affect expression framework?

Studies like [5, 13, 17, 30, 35] have inspected various motion and body languages features. We will focus on some of the recurring features from these studies which have been reported to be effective. Motion features: speed, amplitude, repetition and body language features: vertical head pose, bend-straight stance seems promising and will be inspected. The perceived valence and arousal ratings would indicate the effectiveness of these features and associated operators in expressing specified valence and arousal values.

Many of the above studies focus on limited affects, usually one affect from each quadrant of the valence-arousal plane. Hence, these features may not be adequate for expressing affects mapped onto the same quadrants. For example, fear and anger fall into the same quadrant (negative valence, high arousal). Hence, similar feature modulations are applied in both cases and the resulting affects may not be distinguishable. Some studies have shown that certain affects are easier to recognise than others. [31] found that affects like happy (positive valence, high arousal) and sad (negative valence, low arousal) are more easily recognised than affects like anger (negative valence, high arousal), which have opposite signs for valence and arousal. Additionally, some affects which have comparable arousal values but conflicting valence values, are often confused [5, 30, 35].

It is desirable to develop a model which modifies how a gesture is performed. Unlike emotion-specific pose repertoires, a parametric model tries to cover the entire affect space. Considering the above observations and the constraints, it can be assumed that a parametric model based solely on body language and motion features would have a limited range of perceivable affects. This leads us to the research question:

[Evaluation] 1(b). Which affects expressed by the motion and body language model are recognisable and distinguishable ?

Hypothesis: Happy, sad and excited would be distinguishable and recognisable.

As noted in [3, 31, 35] affects like happy, sad and excited are reliably expressed by modulating motion and body language features. Hence at least these three affects are expected to be recognised. Table 1.2 shows the expected recognition rates of various affects. The valence and arousal ratings of these affects are expected to be close to the intended values and hence distinguishable.

Affect	Motion and body language
Happy	High
Excited	High
Anger	Low
Fear	Low
Sad	High
Tired	Low
Relaxed	Low
Content	Low

Table 1.2: Expected recognition rate of various affects expressed using only motion and body language features. The cells are coloured green for recognition rates ≥ 75%.

Some previous works have studied the relationship between colours and emotions. Many humanoid robots are capable of displaying colours through LEDs on their head, chest or around sensors like camera, microphone, etc. This thesis focuses on using the 'eye LEDs' as an additional channel to express affect. Studies like [14, 28] have demonstrated that some LED colours and blinking patterns in a robot are perceived as specific emotions. These studies focused solely on LED channels and could successfully express a good range of emotions. Hence a model that combines motion and body language model with LED patterns could improve the range of perceived affects. For example, red LED patterns are often associated with anger. As discussed before, anger expressed through motion and body language features is often perceived as happy or excited. Thus, the addition of red LED patterns could improve the recognition of anger. This is the basis for the following research questions:

[Design] 2(a). What are the colours and patterns which can be used for expressing various affects?

[10, 19, 28] have shown that hues of red are associated with high arousal whereas hues of blue are associated with low arousal. [14, 28] have demonstrated that blinking frequency is associated with the arousal of the affects. Hence, hues of red with high blinking frequency can be used to express high arousal affects. Similarly, hues of blue with low blinking frequency can be used to express low arousal affects. [10, 19] noted that green is associated with positive valence affects like relaxed, calm, etc. Hence, the valence ratings of relaxed and content are expected to improve by using LED patterns.

[Evaluation] 2(b). What additional affects are perceived by incorporating LED patterns?

Hypothesis: Anger would be distinguishable and recognisable by the addition of LED patterns.

As observed in [14, 28], anger is reliably expressed through red LED patterns. Hence, anger is expected to be recognised by incorporating LED patterns. The valence and arousal ratings would also reflect this. Due to the association of green with positive valence [10, 19], the recognition rate of relaxed may improve. Emotions like fear would still be a challenge. Table 1.3 shows the expected recognition rates of combining motion and body language fea-

tures and LED patterns.

Affect	Motion and body language	+ LED patterns
Happy	High	-
Excited	High	-
Anger	Low	High
Fear	Low	Low
Sad	High	-
Tired	Low	Medium
Relaxed	Low	Medium
Content	Low	Low

Table 1.3: Expected recognition rate of various affects. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50%−75%. Second column corresponds to the expectations of using only motion and body language features. The third column corresponds to the expected result of including LED patterns.

Like motion and body language features, some affects are easily recognised by the addition of LED patterns. As discussed in [14], anger and happiness are perceived correctly while fear and disgust have low recognition rates. Affects like fear have low recognition rates in motion and body language model as well. Hence it is likely that the combined model still has limits to the range of perceivable affects. In such cases, we have to resort to emotion-specific poses. [3, 7, 12] have shown that humanoid robots can mimic key poses which are associated with various emotions. As alluded to before, such poses often utilise multiple joints in arms, head, etc., which would interfere with the task being performed. This leads us to the last research question:

[Design] 3. What are the emotion specific pose repertoires that can be added to increase the perceivable affects?

Hypothesis: Slightly averted gaze with hands covering the eyes is a key pose that can distinctly express fear.

Since pose repertoires may interfere with the task, this technique is used only for affects which are otherwise not perceived well. Fear could be one such candidate and has easily recognisable key pose involving averted gaze with hands covering the eyes [3, 7]. However, content is an affect which may not be perceived correctly in other models and does not have a well-known key pose. Hence, the expression of content might still be a challenge. Table 1.4 shows the expected improvements after adding pose repertoires for certain affects.

Affect	Motion and body language	+ LED patterns	Pose repertoires
Happy	High	-	-
Excited	High	-	-
Anger	Low	High	-
Fear	Low	Low	High
Sad	High	-	-
Tired	Low	Medium	-
Relaxed	Low	Medium	-
Content	Low	Low	-

Table 1.4: Expected recognition rates of various affects. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50%−75%. Second column corresponds to the expectations of using only motion and body language features. The third column corresponds to the expected results of combining motion and body language model and LED patterns. The last column shows the expected expressiveness by using pose repertoires for selected affects.

Answering these questions gives an insight to the varying complexity of expressing various

affects. It also is a foundation for building a framework which can render a wide range of affects to any non-verbal task or behaviour.

1.4. Thesis Overview

The rest of this thesis is structured as follows. Chapter 2 discusses the prior works on affect expression, focusing on their approach and various features employed by each of them. Chapter 3 explains the kinematics of the social robot NAO, which was used in all the experiments conducted as a part of this thesis. Chapter 4 formulates the mathematical models which form the foundations of the framework. It also elaborates the definitions and modulations associated with the various features. The details about implementing and instantiating the models are described in chapter 5. Chapter 6 describes the gestures involved and the approach followed in the experiments. The experiments were conducted in three phases. The results of phase 1 are presented in chapter 7. The results of the other two phases are presented in chapter 8. Finally, chapter 9 discusses the overall results and highlights the contributions, limitations and future prospects of this work.

2

Related work

Expressing affect facilitates natural and human-like interactions between humans and robots. This is especially important for social robots which are designed to provide interactive services or companionship. Numerous works have studied affect expression in robots which have varying capabilities: from simple arm-robots to robots with facial features. The scope of this thesis is limited to humanoid robots with minimal body features. This chapter gives a brief overview of some related works which studied affect expression in robots.

2.1. Motion and Body-language based models

Motion is inevitable for a robot while performing tasks. Modulating the motion and body language features to express affect has gathered a lot of attention because such a model can be applied to many robots.

[16] proposes a Laban efforts based framework to generate expressive motions. This framework modifies various motion features like velocity, acceleration, abruptness and arrival time, to convey the robot's attitude. The experiments involved 2 robots (NAO and Keepon) and 2 tasks (look-around and dance). Along with the motion features, body language features like vertical compression, head pose, etc. were modulated to generate expressive motion.

The paper [21] proposes a Laban shape and efforts based framework for expressing affect. Laban efforts were modelled using speed, smoothness, duration and frequency of motion. Approach/avoidance was modelled using Laban shape features like leaning forward vs backwards, expand vs shrink, etc. Angry, fear, happy, sad and surprised versions of the gestures were analysed. It was found that speed and smoothness indicated arousal, whereas duration and frequency influenced valence, arousal and dominance. Approach/avoidance portrayed valence and dominance. Participants could recognise angry, fear, happy and surprised, but not sad.

[9] studied the Laban effort and shape profiles in walking gesture. This model used features similar to [16, 21]. The participants were presented various versions of walking portraying anger, joy, content and sad. Sad had high recognition rate. Joy and anger were often confused and neutral was mostly recognised as content.

The authors of [2] proposed modulation of amplitude and speed to generate emotional gestures. The study involved 2 gestures (drinking and kicking) and 2 emotions (sad and angry). First, the angry and sad transformations were calculated from an actor's portrayal of emotional drinking. These transformations were then applied to the neutral gestures to generate the emotional versions. The generated emotional gestures were comparable to the original enactments. The paper also examined the frequency of joint positions as an addi-

tional motion feature. However, this feature did not improve the results for either emotion.

[17] proposes emotion-specific features to emulate angry, happy and sad walking gestures. The study was conducted on a bipedal humanoid robot (WABIAN-RII). The features included speed, step-length and bending forward vs backward. Happy and sad had high recognition rates. Due to the joint constraints of the robot, the complete sequence for angry walking was not executed. Hence, anger had slightly low recognition rate.

The paper [35] proposes modulation of motion features through adjectival words. The motion features considered were speed, amplitude, display position and acceleration. This approach calculates the correlations of emotions with adjectival words like wide, slow, low, etc. Each of the adjectival words was associated with a pre-defined feature modification. An emotional gesture was generated by applying modifications corresponding to all the correlated adjectival words. For example, joy (associated with wide and fast) results in 2 times the amplitude and 1.5 times the speed. This model was tested on 4 gestures (wave 1, wave 2, greeting and handing over) and 4 emotions (joy, sad, angry, fear). Analyses revealed that acceleration was an insignificant feature and speed was the most important feature. Joy and anger were often confused because they both are associated with the adjective: fast. The paper suggests using different levels of fast to differentiate them.

In [5], the authors propose a model based on motion features to express 4 emotions (angry, sad, joy, pleasure), belonging to the four quadrants of valence-arousal space. The study examined affective videos and extracted motion features like velocity, acceleration, area of motion, fluidity and contraction/expansion. The results show that the area of motion was the most significant indicator of arousal, while contraction/expansion portrays negative and positive valences. In experiments employing this model, anger had a high recognition rate, whereas sad, joy and pleasure had moderate recognition rates. Additionally, negative and positive emotions with the same arousal levels were often confused.

[13] asked the participants to choose a head pose associated with 6 emotions (angry, fear, disgust, sad, happy, surprise). Happy and surprise were associated with a head-up pose, whereas angry and sad were associated with a head-down pose. Though disgust and fear were on an average associated with an averted or look-away pose, the responses had a large variance.

[30] proposes a parametric mood expression model based on an extensive list of motion and body language features: speed, amplitude, hand-height, palm up-down, finger rigidness, decay speed, hold-time, repetition, horizontal and vertical head poses. Out of this initial list, repetition, decay speed and hold-time were considered as indicators of arousal, and others as indicators of valence. This model could express happy, excited and sad on 2 gestures (wave and pointing). These features were further studied in [31, 32, 34]. Amplitude, speed, hand-height, repetition and vertical head pose were relatively more significant in expressing affect [31, 32]. [34] attempted to classify these features as valence-oriented or arousal-oriented. Speed and repetition were classified as arousal-oriented features. Amplitude influenced both valence and arousal but had more affinity towards valence. [33] focused on expressing very sad, sad, neutral, happy and very happy by modulating these features. These affects were expressed using only the valence values.

The studies mentioned in this section have demonstrated that motion and body language features can be used to express affects like sad, happy, etc. However, most of these studies formulate modulations for specific affects. [30] formulated a parametric model which focused on a valence-oriented mood expression. This thesis uses this approach to build a 2D affect expression framework. The features which appear in multiple studies are chosen for building the motion and body language model. Out of the recurring features, the definition of amplitude varies from paper to paper. [30] used gesture-specific amplitude definitions. [16] only modified the amplitude of yaw angles, which works for simple gestures like look-around,

but not for other gestures like waving and handshake. This thesis uses a generic definition adapted from [2].

2.2. Colour based models

Several works have studied the relationship between colour and emotions. Humanoid robots mostly have LEDs on their bodies, typically near the 'eye'. These LEDs are generally used to communicate coloured error codes indicating the robot's state. However, they are seldom used and thus can serve as an additional channel for expressing affect. Some of the works which studied the relationship between colour and emotion expression are discussed below.

[19] studied the emotions associated with hues of red, yellow, green, blue and purple. The authors noted that colours were often symbolic, and people tend to associate them with specific concepts or memories. Green was associated with positive emotions like relaxed, calm, restful, etc., and yellow or yellow-red was associated with happiness, excitement and joy. Hues of blue represented low arousal emotions while red signified high arousal emotions like energetic, angry, love etc.

The framework proposed in [29] modulates features like eye colour, body pose, speech volume, speech rate and gesture size to express affect. It employs multiple channels rather than focusing on a single channel. The speech parameters, eye colours and gesture size were modulated by arousal, whereas body pose was influenced by both valence and arousal. The LEDs displayed hues of red for high arousal emotions and hues of blue for low arousal emotions.

In [14], the authors studied colours and blinking patterns associated with 6 emotions (surprise, happy, sad, disgust, anger, fear). Red portrayed anger, whereas yellow depicted happy and surprise. Disgust was associated with dark-green, sad with blue or cyan, and fear with grey. The blink pattern starts with no colour, rises in the intensity of colour and then falls back to no colour. This pattern was repeated periodically to emulate blinking. Anger was the only emotion which was recognised using these patterns. The paper further investigated several patterns which imitate cartoonish facial expressions. For example, a small part of the eye LED was coloured blue to portray tears while crying, glowing red eyes for anger, circling bright colours for happiness, etc. Anger, happy, sad and surprise had high recognition rates when using the cartoonish patterns. However, fear and disgust were often recognised as sad.

[10] proposes an affective messaging model for mobile phones. It varied the colour and size of objects in the background for expressing affect. The paper proposed mapping a hue circle onto the valence-arousal plane. The high arousal emotions were associated with hues of red. Emotions like calm, relaxed, serene, etc. were associated with hues of green. Emotions in the third quadrant like sad, tired, miserable etc. were associated with hues of blue.

[28] proposes a model to express Plutchik's 8 basic emotions (anger, anticipation, joy, trust, fear, surprise, sadness, disgust) and their variations. It proposed hue values and blinking patterns for expressing various affects. The frequency of blinking varied depending on the arousal of the emotion. For example, high arousal emotions like anger and amazement had high blink rates, whereas low arousal emotions like serenity and boredom had low blink rates. Arousal also determined the shape of the blink waveform, which indicated the smoothness of intensity variations. High arousal emotions had high-frequency square waveforms, and low arousal emotions had low-frequency bell-shaped waveforms. This model succeeded in expressing Plutchik's 8 basic emotions. Except rage, vigilance, amazement and loathing, the variations of basic emotions were also expressed.

The studies in this section focused on colours associated with affects. Among them, [14, 28, 29] conducted experiments on robots. [14, 28] focused on expressing specific emotions. This thesis adopted the hue-circle model proposed in [10] to develop a parametric model for

modulating LED features. [29] also proposes a parametric model which determines the LED colour based on the arousal. Such a model may yield similar colour for affects which differ mainly on valence, e.g. sad and content.

2.3. Pose based models

This is one of the easier techniques to express emotions. These models use an emotion-specific pose to depict an emotion. One of the drawbacks of such models is that adding support for a new emotion requires designing a new key pose. The following studies explored pose based models.

The authors of [3] designed unique NAO robot poses for 6 emotions (anger, sad, fear, pride, happy, excited). These poses were modelled using the motion-capture and video recordings of professional actors who enacted emotion-specific poses. Sad was portrayed using bend knees, leaning forward and head down. Fear was depicted using a backward stance and hands partially covering the face. The poses used in this paper can be seen in figure 2.1. All the poses had recognition rates $\geq 73\%$.

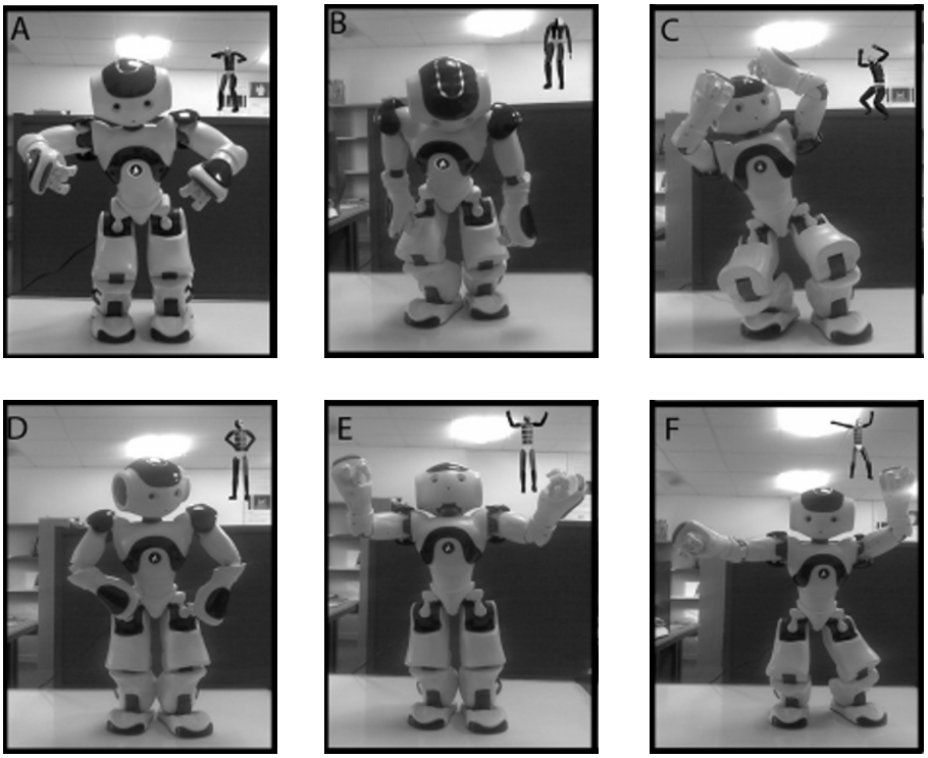


Figure 2.1: The key poses taken from [3]. The emotions expressed are A: Anger, B: Sad, C: Fear, D: Pride, E: Happy, F: Excited

.

In [7], the authors modelled 5 emotion-specific poses (anger, fear, happy, surprised, sad) of NAO. Again fear was depicted by covering the face with a hand. There are similarities in happy and sad poses portrayed in this paper and [3]. The 5 key poses can be found in figure 2.2. Some poses were recognised more than others. The recognition rates of all the poses were more than 67%.

Studies using pose repertoires focus on expressing specific emotions. This implies that this technique cannot be used in a parametric model. This thesis studies the incremental expressive capability of each of the techniques. Pose repertoires are only employed when the previous models are not sufficient for expressing certain affects. Such affects are seen as complex affects, i.e. they are hard to express using simple features and parametric modulations.

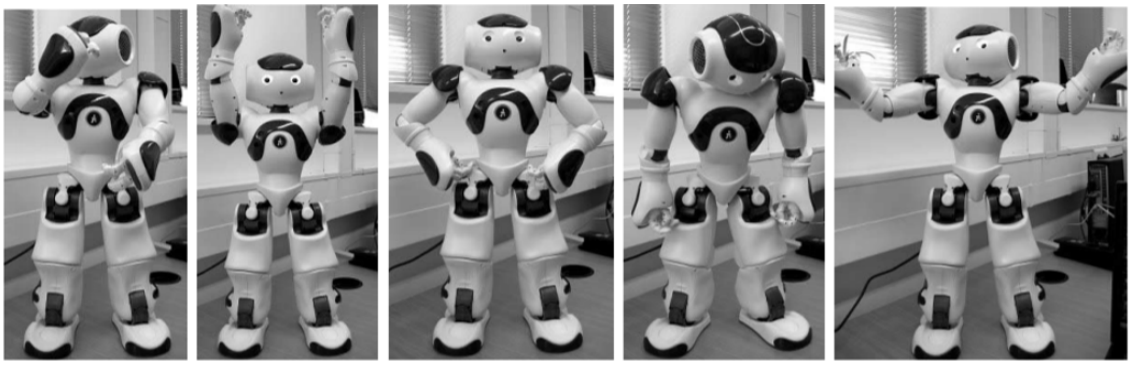


Figure 2.2: The key poses taken from [7]. The emotions expressed are (from left to right) fear, happy, angry, sad and surprised

.

2.4. Facial expression based models

These models are designed for robots which are capable of imitating the facial expressions of humans. Controllable facial features like eyebrows, lips, eyelids etc. are important for such models. This thesis focuses on simple humanoid robots and does not employ facial expressions to portray affect. Since facial expressions also qualify as a non-verbal technique for expressing affect, we include a brief discussion of a couple of papers which employed this technique.

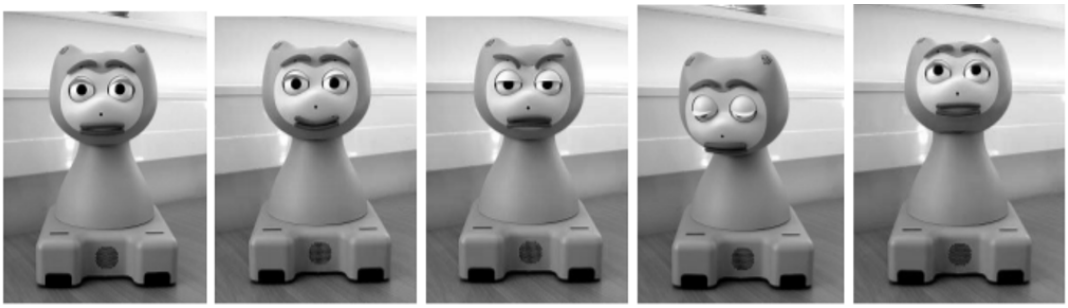


Figure 2.3: The facial expressions of iCat robot presented in [7]. The emotions expressed are (from left to right) fear, happy, angry, sad and surprised

.

The authors of [7] designed facial expression for an iCat robot. It expressed 5 emotions which were same as the ones used for studying emotion-specific poses. Figure 2.3 shows the various facial expressions designed for iCat. The recognition rates of the emotions were at least 69%.

Figure 2.4: The facial expression of Robotinho developed in [20]. The emotions expressed are (from left to right) joy, surprise, anger, sad, disgust and fear

.

[20] proposes a multi-modal emotion expression framework for the Robotinho robot. Along with the speech features, facial features illustrated in figure 2.4 were employed to improve the interaction capabilities of the robot. As a part of the study, the robot was deployed as

a museum tour guide. The interactions were rated as friendly and intuitive by adults and children.

3

Development platform

This thesis focuses on affect expression in simple humanoid robots. NAO is a bipedal interactive humanoid robot from SoftBank Robotics[1] which was first launched in 2006. NAO has a head with 2 DOF (Degrees Of Freedom) and arms with 6 DOF each, which can perform human-like movements. The experiments in this thesis use NAO V6, which was launched in 2018. NAO is widely used in studies focusing on social robots and their human interactions. Due to its small size, it is quite popular in research involving robot interactions with children.

3.1. Technical Overview

NAO V6 has a height of 57.4 cm and weighs 5.5 kg. NAO works on battery charge as well as when plugged to a power source. It can communicate via an IEEE 802.11g wireless or a wired Ethernet port located behind its head. It has peripherals such as loudspeakers, LEDs, microphones and video camera to facilitate interaction.

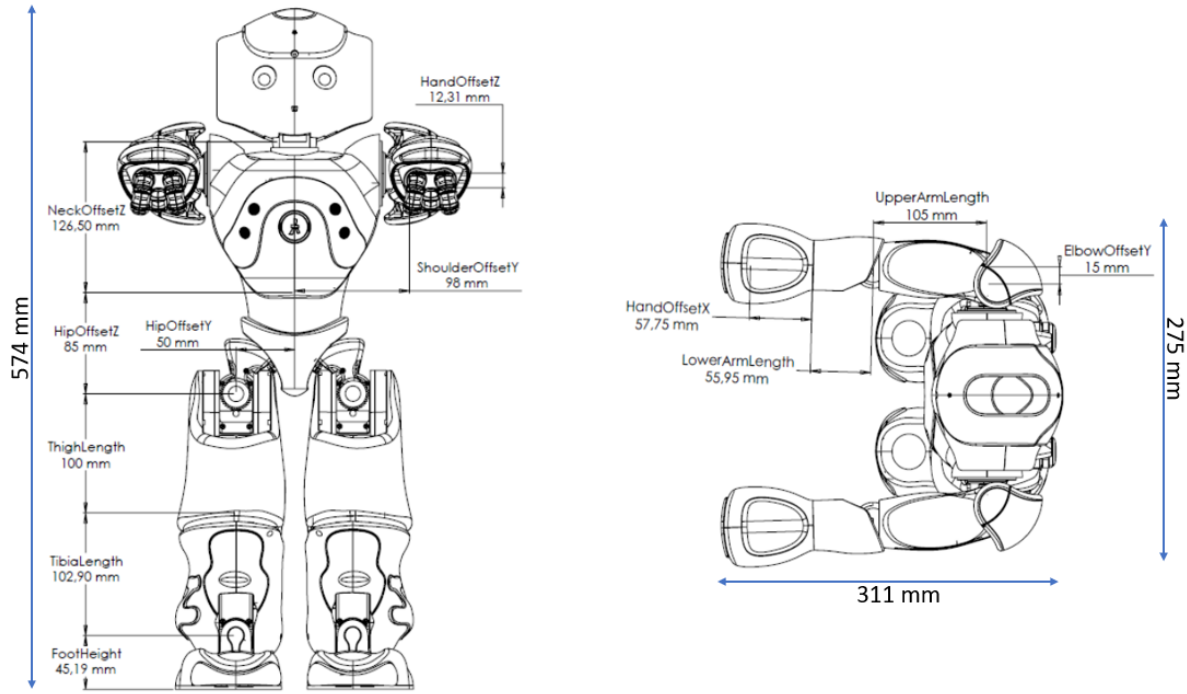


Figure 3.1: The height, width and depth details of NAO, obtained from Aldebaran website

[1] https://www.softbankrobotics.com/emea/en/nao

3.2. Joints

The joint motions are defined as rotations along X, Y and Z axes. The convention followed is that the X-axis is NAO's back to front, the Y-axis from right to left and the Z-axis is vertical. As seen from figure 3.2, roll rotations are around the X-axis, pitch rotations around the Y-axis and yaw rotations around the Z-axis.

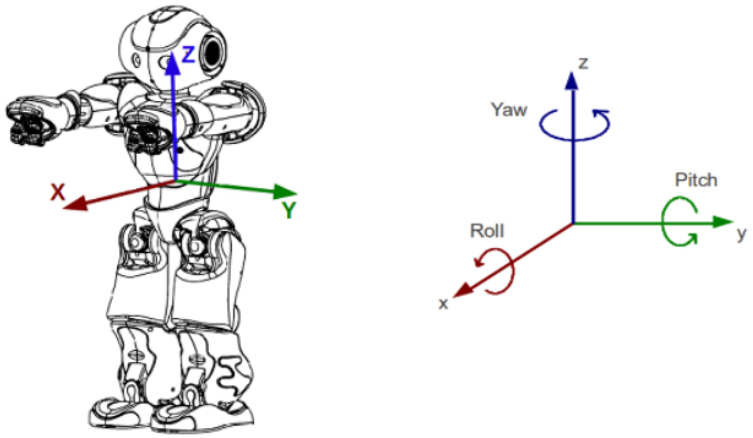


Figure 3.2: The axis conventions followed for NAO, obtained from Aldebaran website

Inspired by human anatomy, NAO has five joint chains corresponding to Head, Left Arm, Right Arm, Left Leg and Right Leg. Figure 3.3, shows the joint chains and the joints belonging to each of them. Head has two joints - yaw and pitch. Each arm has six motors controlling the joints - shoulder pitch, shoulder roll, elbow yaw, elbow roll and wrist yaw. Each leg has six joints - hip yaw-pitch, hip roll, hip pitch, knee pitch, ankle pitch and ankle roll.

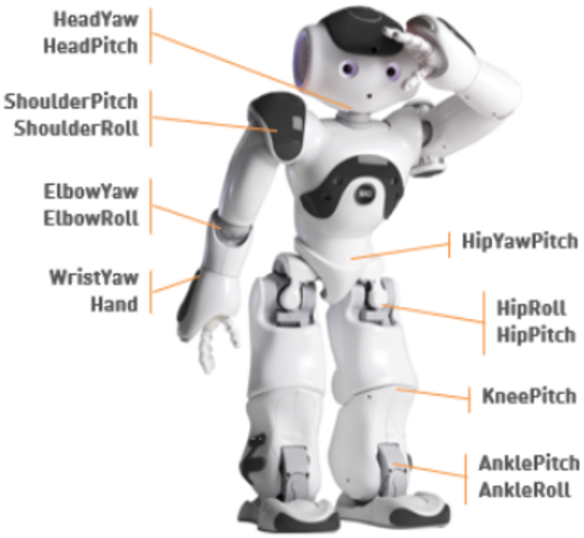


Figure 3.3: Illustration of NAO robot showing various joints, obtained from Aldebaran website.

3.3. Motions and Joint Constraints

Each joint has different motions associated with them. As seen in figure 3.4, NAO can perform two kinds of head motions: up-down and left-right motions.

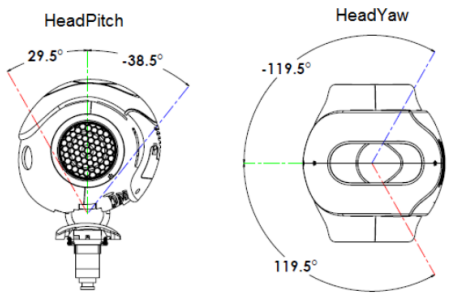


Figure 3.4: Joints and range of motion for head taken from Aldebaran website

Figure 3.5 illustrates the joints and the associated range of motions for the left arm. The right arm has similar joints and range of motions as the left arm. The hand joint in the arms control the openness of fingers, where 1 represents wide open fingers and 0 represents a closed configuration. The range of motions associated with the hip and left leg are illustrated in figure 3.6. The ranges are mirrored for the right leg. Though hip yaw-pitch has left and right joint labels, the same motor controls both these joints. In case of a disparity between the left and right values, only the left hip yaw-pitch value is regarded.

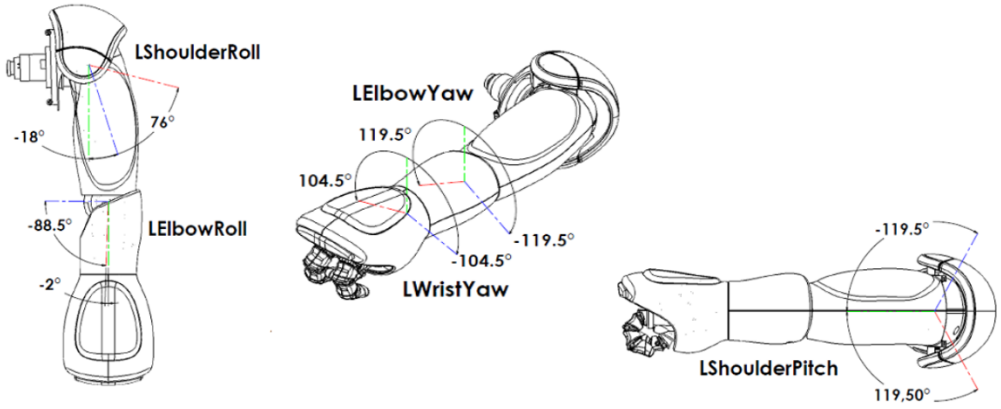


Figure 3.5: Joints and range of motion for left arm, taken from Aldebaran website

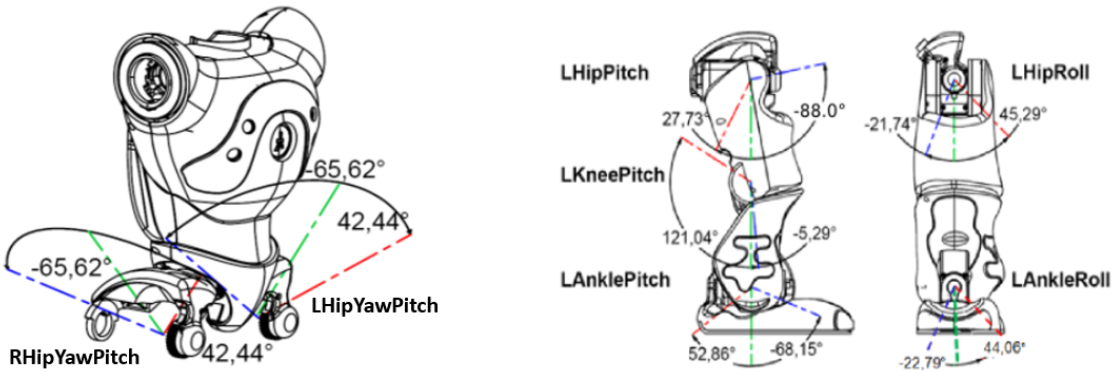


Figure 3.6: Joints and range of motion for hip and left leg, taken from Aldebaran website

As seen in the figures above, there are limits to the range of motion or the angle a joint can sweep, and it differs from joint to joint. Joints that exist on both left and right joint chains have mirrored ranges. These ranges act as constraints that need to be taken into account

Joint name	Range (degrees)	Range (radians)
HeadYaw	-119.5 to 119.5	-2.086 to 2.086
HeadPitch	-38.5 to 29.5	-0.672 to 0.515
LShoulderPitch	-119.5 to 119.5	-2.086 to 2.086
LShoulderRoll	-18 to 76	-0.314 to 1.327
LElbowYaw	-119.5 to 119.5	-2.086 to 2.086
LElbowRoll	-88.5 to -2	-1.545 to -0.035
LWristYaw	-104.5 to 104.5	-1.824 to 1.824
RShoulderPitch	-119.5 to 119.5	-2.086 to 2.086
RShoulderRoll	-76 to 18	-1.327 to 0.314
RElbowYaw	-119.5 to 119.5	-2.086 to 2.086
RElbowRoll	2 to 88.5	0.035 to 1.545
RWristYaw	-104.5 to 104.5	-1.824 to 1.824
LHipYawPitch	-65.62 to 42.44	-1.145 to 0.741
LHipRoll	-21.74 to 45.29	-0.38 to 0.791
LHipPitch	-88.00 to 27.73	-1.536 to 0.484
LKneePitch	-5.29 to 121.04	-0.092 to 2.113
LAnklePitch	-68.15 to 52.86	-1.19 to 0.923
LAnkleRoll	-22.79 to 44.06	-0.399 to 0.769
RHipRoll	-45.29 to 21.74	-0.791 to 0.38
RHipPitch	-88.00 to 27.73	-1.536 to 0.484
RKneePitch	-5.29 to 121.04	-0.092 to 2.113
RAnklePitch	-67.97 to 53.40	-1.187 to 0.932
RAnkleRoll	-44.06 to 22.80	-0.77 to 0.398

Table 3.1: Joints and the constraints on their range of motion, obtained from Aldebaran website. The radians are rounded to three decimal places

while designing or modulating a gesture. The range of motion for the joints in the head, arms and legs are given in table 3.1

3.4. LEDs

NAO has multiple LEDs on its head, eyes, ears and feet. While head and ears only display various levels of white and blue respectively, eyes and feet can display a full range of RGB colours. Each foot has only one LED. In this work, only the eye LEDs are used to express affect. As illustrated in figure 3.7, there are 8 LED spots in each eye spaced uniformly at 45 degrees. Each spot has 3 LEDs, one for each colour channel of RGB. Every LED has an intensity value ranging from 0.0 (no light) to 1.0 (full light), which can be manipulated to output a wide range of RGB colours. The colours of eye LEDs can be set at Face/Led/[Colour]/[Side]/[Degree], where [Colour] can be Red, Green or Blue, [Side] can be Left or Right and [Degree] can be 0Deg, 45Deg, 90Deg, 135Deg, 180Deg, 225Deg, 270Deg or 315Deg.

3.5. NAOqi APIs

NAOqi is an SDK provided by SoftBank robotics for programming their social robots like NAO and Pepper. The NAOqi APIs are compatible with multiple programming languages like python, C++, Java, etc. Currently, only NAOqi version 2.8 supports NAO V6.

The **ALMotion** API is used for making NAO perform any motion. It is responsible for joint stiffness (switching the motor on/off) and joint positioning (used for generating movement). This API expects a list of angles and associated timestamps for each joint, which can be interpolated to generate gestures. Joints can be interpolated individually or simultaneously as a group.

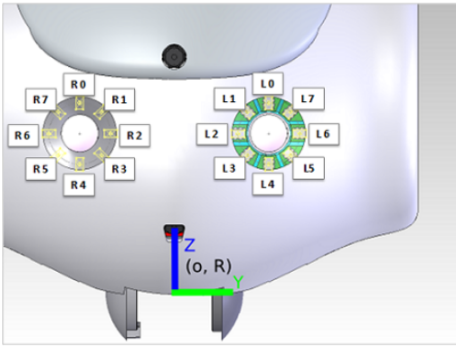


Figure 3.7: LED positions on NAO's eyes, taken from Aldebaran website

The LEDs are controlled through **ALLeds** API. For any given LED, this API expects a colour and the time to achieve the desired colour. Once set, the LED retains the colour until it is reset or the robot is powered off. The API gives the option of controlling each LED separately or as a group of LEDs.

4

Features and modulations

This chapter introduces and defines explicitly, the motion, body language and LED features involved in expressing affects. It also defines the associated operators and parameters which enables us to map the modulations onto the hardware of a robot. First, some mathematical representations and notations are introduced to formulate the models for the various features. The various motion, body language and LED features and the operators for modulating these features are formulated in the subsequent sections.

4.1. Gesture Representation and Notation

There are many ways to represent a gesture which serves as the input to the framework and the models. This thesis uses an *angle-time* representation where a gesture is defined as a *time series of poses*. A *pose* is a vector of the joint angles of the robot.

A *(complete) pose vector* consists of all the joint angles of a robot and represents a full pose of the robot. In the case of NAO, a pose vector $\boldsymbol{\theta}$ involves angles listed in table 3.1 and would look like:

$$\boldsymbol{\theta} = \langle \theta^{HeadPitch}, \theta^{HeadYaw}, \dots, \theta^{LeftElbowYaw}, \theta^{RightHipRoll}, \dots \rangle$$

$\boldsymbol{\theta}^i$ denotes the ith component of the vector $\boldsymbol{\theta}$. The components of a vector can also be written using specific joint names. For example, $\boldsymbol{\theta}^{HeadYaw}$ denotes the $HeadYaw$ angle in vector $\boldsymbol{\theta}$. A *partial pose (vector)* $\boldsymbol{\theta}$ is a pose vector where some of the angle values are set to ϵ (denotes missing or null value), e.g. $\boldsymbol{\theta} = \langle \epsilon, \epsilon, \dots, \theta^{LeftElbowYaw}, \dots \rangle$. Two partial poses can be added together, $\boldsymbol{\theta} = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2$, as follows:

$$\boldsymbol{\theta}^i = \begin{cases} \boldsymbol{\theta}_2^i & \text{if } \boldsymbol{\theta}_1^i = \epsilon, \\ \boldsymbol{\theta}_1^i & \text{otherwise} \end{cases}$$

This plus operator is *not* symmetric. If a joint angle is specified in both pose vectors, the resulting vector always has the angle value of $\boldsymbol{\theta}_1$.

A *marked pose*, written as $\boldsymbol{\theta}^*$, indicates that $\boldsymbol{\theta}$ is a so-called *pivot point* (or simply *pivot*) in a gesture, i.e. a time series of poses. The amplitude of pivot points should not be modified to avoid changing the nature of the gesture (see section 4.2). $\boldsymbol{\Theta}$ represents the space of all *poses*, i.e. all partial and complete, and possibly, marked poses $\boldsymbol{\theta}$ that a robot can achieve.

A *timed pose* is a pair $(\boldsymbol{\theta}, t)$ with $t \in \mathbb{R}^+$ $(0 \notin \mathbb{R}^+)$. The pose vector and timestamp can be extracted from a timed pose using π_1 and π_2 operators as follows: $\pi_1(\boldsymbol{\theta}, t) = \boldsymbol{\theta}$ and $\pi_2(\boldsymbol{\theta}, t) = t$. A *gesture* or *motion* δ is a sequence of timed poses $\delta = \langle (\boldsymbol{\theta}_1, t_1), \dots, (\boldsymbol{\theta}_n, t_n) \rangle$ such that $t_{i+1} > t_i$. In *well-defined* gestures, the robot can achieve $\boldsymbol{\theta}_1$ in time t_1 and interpolate all subsequent poses $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_{i+1}$ in the time span of $t_{i+1} - t_i$ for all $1 \leq i < n$. δ^i denotes the timed pose $(\boldsymbol{\theta}_i, t_i)$

in gesture δ. $\mathbf{\Delta}$ represents the space of all gestures or motions that a robot can perform.

The amplitude operator modulates the amplitude of poses lying between two pivot points while keeping the pivot points intact (see section 4.2). A pose $\pi_1(\delta^j) = \boldsymbol{\theta}$ in a gesture δ *lies between pivot points* $\boldsymbol{\theta^*}_{before}$ *and* $\boldsymbol{\theta^*}_{after}$ if $\boldsymbol{\theta}$ is not itself a pivot point, $\pi_1(\delta^i) = \boldsymbol{\theta^*}_{before}$ is a pose before $\boldsymbol{\theta}$, i.e. $i < j$, and $\pi_1(\delta^k) = \boldsymbol{\theta^*}_{after}$ is a pose after $\boldsymbol{\theta}$, i.e. $j < k$, and there are no pivot points $\pi_1(\delta^l)$ with $i < l < j$ or $j < l < k$. Additionally, a pose $\boldsymbol{\theta}$ in a gesture δ is said to *lie between pivot points* $\boldsymbol{\theta^*}_{before}$ *and* $\boldsymbol{\theta^*}_{after}$ if $\boldsymbol{\theta}$ is not itself a pivot point and δ only has a single pivot point $\boldsymbol{\theta}'$ and $\boldsymbol{\theta^*}_{before} = \boldsymbol{\theta}'$ and $\boldsymbol{\theta^*}_{after} = \boldsymbol{\theta}'$.

δ^{+t} denotes the gesture where all times associated with poses have been increased by t, i.e. for a gesture δ of length k, $\pi_2((\delta^{+t})^i) = \pi_2(\delta^i) + t$ for all $1 \leq i \leq k$. $\delta_{i,j}$ with $0 < i,j < k$ denotes the (sub)motion $\langle(\boldsymbol{\theta}_i, t_i), \ldots, (\boldsymbol{\theta}_j, t_j)\rangle$ for $\delta = \langle(\boldsymbol{\theta}_1, t_1), \ldots, (\boldsymbol{\theta}_k, t_k)\rangle$. The concatenation of two gestures can be written as $\delta_1 + \delta_2$, where times have been increased in the second motion by the final time t_k in δ_1, i.e., $\delta = \delta_1 + \delta_2 = \delta_1 \cdot \delta_2^{+t_k}$ where $\cdot$ represents the usual concatenation of two sequences, δ_1 has length k, and $\pi_2(\delta_1^k) = t_k$. Note that $\pi_2(\delta^k) \neq \pi_2(\delta^{k+1})$ as $\pi_2(\delta_2^1) > 0$.

As a simple example, consider figure 4.1 in which the robot interpolates 3 different poses to generate a wave gesture. Formally, it can be represented as: $\delta_{wave} = \langle(\boldsymbol{\theta^*}_1, t_1), (\boldsymbol{\theta}_2, t_2), (\boldsymbol{\theta}_3, t_3)\rangle$. Note that the first pose in this gesture has been marked as a pivot point.

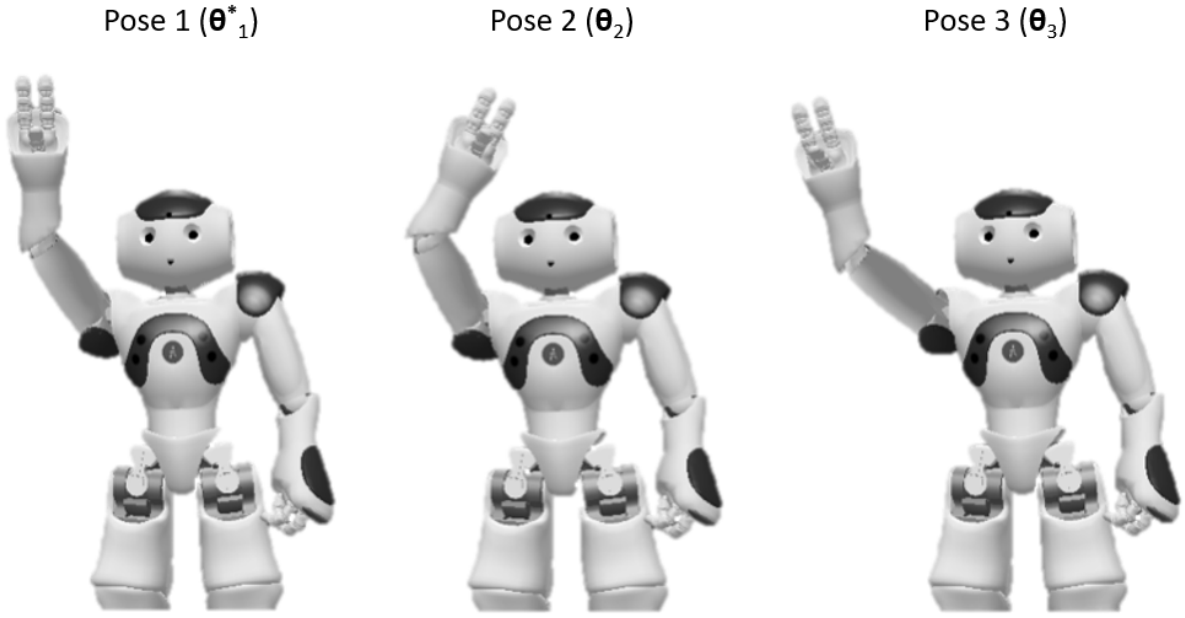


Figure 4.1: Illustration of a simple waving motion broken down to 3 poses. The initial pose or $\boldsymbol{\theta^*}_1$ is a pivot point where the right arm is raised up in a central position. The right arm first moves inwards to pose 2 ($\boldsymbol{\theta}_2$) and then outwards to pose 3 ($\boldsymbol{\theta}_3$).

4.2. Motion and Body language operators

An essential step in generating affective gestures is finding a minimal set of features and operators to modulate these features. As discussed in previous chapters, [31, 32] modulated a small set of motion and body language features to generate affective gestures. [9, 16, 21, 25] modulate Laban components to portray affect. There is a significant overlap between features used in these Laban models and the parametric models in [5, 30–32]. As seen in chapter 2, several studies have demonstrated that features like speed, amplitude, head pose, etc. are very significant in portraying affects. These features pertain to movement or body language and are not robot-specific. Some of these recurring features and the associated modulation operators are defined below.

1. Amplitude: A gesture has pivot points (or fixed points) that should not be altered while modulating amplitude. In the case of pointing, change in the final configuration of angles would change the direction of pointing. Thus, the final pointing pose is a pivot point. In case of a wave gesture, an increase in amplitude increases the angle covered by the hand from the central position as seen in figure 4.2. So this centre pose is the pivot point for waving.

   Given a time series of joint angles, the *amplitude* of a point is its distance from a *reference line*. As illustrated in figure 4.3, the reference line is a line between two consecutive pivot points. Amplitude is modified for each joint and independent of other joints.

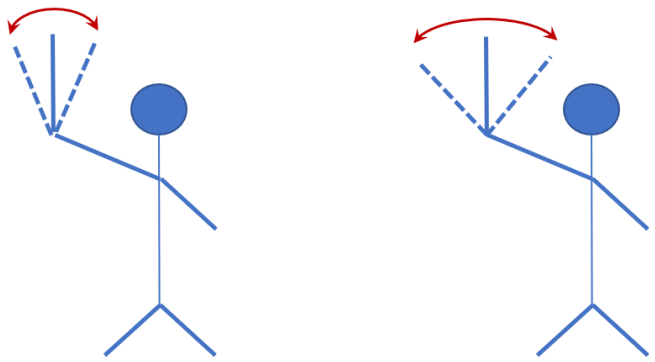


Figure 4.2: The first stick figure depicts a normal waving and the second shows waving with increased amplitude. The extreme hand positions are changed to increase the angle covered during the motion while the central position is unaltered.

   The algorithm for increasing or decreasing the amplitude, as proposed in [2], is based on the observation that the points near the pivot points have smaller changes in amplitude compared to the points away from the pivot points. Consider a motion which has two pivot points at say, $(\boldsymbol{\theta^*}_1, t_1)$ and $(\boldsymbol{\theta^*}_n, t_n)$. For any joint j, a reference line can be drawn between $(\boldsymbol{\theta}_1^j, t_1)$ and $(\boldsymbol{\theta}_n^j, t_n)$ and the distance of any point $(\boldsymbol{\theta}_i^j, t_i)$ from this line is the amplitude at that point. As illustrated in figure 4.3, the amplitude can be modified by changing the distance of the point from the reference line. In cases where there is only one pivot point $(\boldsymbol{\theta^*}_p, t_p)$, the reference line is drawn parallel to the time axis passing through $\boldsymbol{\theta}_p^j$.

   The *amplitude operator* A^α, where α is the amplitude factor, can be applied to a gesture δ of length n by the following constraints, for all $1 \le i \le n$, with $\delta^i = (\boldsymbol{\theta}_i, t_i)$, $t_n = \pi_2(\delta^n)$, and $(\tilde{\boldsymbol{\theta}}_i, \tilde{t}_i) = (A^\alpha(\delta))^i$:

$$\tilde{\boldsymbol{\theta}}_i^j = \begin{cases} \alpha \times \boldsymbol{\theta}_i^j + (1-\alpha)(\boldsymbol{\theta}_a^j \times (\frac{t_b - t_i}{t_b - t_a}) + \boldsymbol{\theta}_b^j \times \frac{t_i - t_a}{t_b - t_a}) & \text{if } \boldsymbol{\theta}_i \text{ lies between pivots } \boldsymbol{\theta}_a \text{ and } \boldsymbol{\theta}_b \\ \alpha \times \boldsymbol{\theta}_i^j + (1-\alpha)\boldsymbol{\theta}_p^j & \text{if } \boldsymbol{\theta}_i \text{ lies beyond a terminal pivot } \boldsymbol{\theta}_p \end{cases} \tag{4.1}$$

$$\tilde{t}_i = t_i \tag{4.2}$$

   In Equation 4.1, using $\alpha > 1$ increases the amplitude whereas $\alpha < 1$ decreases the amplitude. Note that the time t_i is normalized to lie within the range $[0, 1]$. Equation 4.2 implies that changing the amplitude does not change the time associated with a pose.

2. Repetitions: Repeating a gesture δ involves going through all the poses from initial pose to end pose in δ, multiple times. Formally, a *repetition operator* $R^k(\delta)$, where k is the number of times gesture δ is repeated, can be defined recursively as follows, for all $k > 0$:

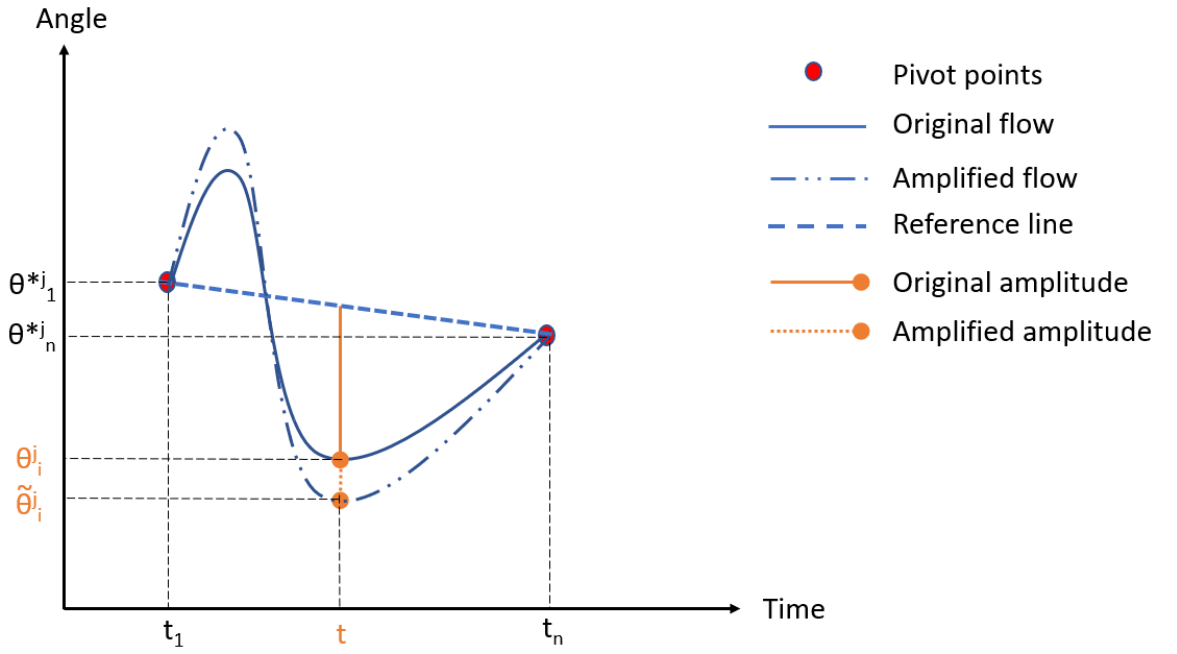


Figure 4.3: An illustration of the amplitude modification algorithm adopted from [2]

$$R^1(\delta) = \delta \tag{4.3}$$
$$R^{k+1}(\delta) = R^k(\delta) + \delta \tag{4.4}$$

3. Speed: Consider a joint movement with angle values $\{x_1, \dots, x_n\}$ at times $\{t_1, \dots, t_n\}$. Speed of the joint in any state i is defined as

$$v_i = \frac{d\,x_i}{d\,t_i} = \frac{x_{i+1} - x_i}{t_{i+1} - t_i}$$

Hence, to modify velocity by a factor of, say α, it is sufficient to divide all t_i by a factor of α. $\alpha > 1$ results in faster movement, and $\alpha < 1$ makes it slower.

$$\frac{x_{i+1} - x_i}{\frac{t_{i+1}}{\alpha} - \frac{t_i}{\alpha}} = \alpha\ v_i$$

Formally, *velocity operator* V^α is defined as follows, for all $1 \le i \le n$, with $\delta^i = (\boldsymbol{\theta}_i, t_i)$, and $(\tilde{\boldsymbol{\theta}}_i, \tilde{t}_i) = (V^\alpha(\delta))^i$:

$$\tilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i \tag{4.5}$$
$$\tilde{t}_i = \frac{t_i}{\alpha} \tag{4.6}$$

4. Head up-down: This thesis uses the vertical head position as a feature which can be modified by controlling the *head pitch*. The ranges of head up and down may differ and hence requires different modification rules. The *vertical head pose operator* $H^{(\alpha,\ flag)}$ takes two parameters: α, the head pitch factor and $flag$, which indicates whether the operation is for *up* or *down* position. A head pose vector $\boldsymbol{\theta}_{hp}(\alpha, flag)$ is a partial pose vector that only specifies the *HeadPitch* angle, i.e. $\boldsymbol{\theta}_{hp}(\alpha, flag)^j = \epsilon$ for all $j \neq HeadPitch$ and

$$\boldsymbol{\theta}_{hp}(\alpha, flag)^{HeadPitch} = \begin{cases} \alpha \times up_{max} & \text{if } flag = up, \\ \alpha \times down_{max} & \text{if } flag = down \end{cases}$$

where up_{max} denotes the maximum angle for head up position and $down_{max}$ denotes the maximum angle for head down position. The operator $H^{(\alpha,flag)}$ is defined as follows, for all $1 \leq i \leq n$, with $\delta^i = (\boldsymbol{\theta}_i, t_i)$ and $(\widetilde{\boldsymbol{\theta}}_i, \tilde{t}_i) = (H^{(\alpha,\ flag)}(\delta))^i$:

$$\widetilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{\theta}_{hp}(\alpha, flag) \quad (4.7)$$

$$\tilde{t}_i = t_i \quad (4.8)$$

5. Stance: This body language feature was inspired by the studies in [9, 21]. This thesis considers 3 stances: neutral, upright/expanded and bend/shrunk, as illustrated in figure 4.4. These stances can be achieved by modifying leg joints L, which for Nao are the joints at hip *(LHipYawPitch, RHipYawPitch, LHipRoll, RHipRoll, LHipPitch, RHipPitch)*, knee *(LKneePitch, RKneePitch)* and angles *(LAnklePitch, RAnklePitch, LAnkleRoll, RAnkleRoll)*. The 3 stances can be represented by the partial pose vectors $\boldsymbol{\theta}_{neutral}$, $\boldsymbol{\theta}_{expand}$ and $\boldsymbol{\theta}_{shrink}$:

$$\boldsymbol{\theta}_s^i(\alpha) = \begin{cases} s_1^i & \text{for } i \in L \text{ and } \alpha \geq \tau_1, \\ s_0^i & \text{for } i \in L \text{ and } \tau_2 < \alpha < \tau_1, \\ s_2^i & \text{for } i \in L \text{ and }, \alpha \leq \tau_2, \\ \epsilon & \text{otherwise} \end{cases}$$

τ_1 and τ_2 denote pre-defined thresholds of α, the parameter that determines the stance to be used. s_0, s_1 and s_2 denote the partial pose vectors which contain valid entries for leg joints in L.

Formally, the *stance operator* S^α which applied to a gesture δ yields a new gesture $S^\alpha(\delta)$, is defined as follows, for all $1 \leq i \leq n$, with $\delta^i = (\boldsymbol{\theta}_i, t_i)$, and $(\widetilde{\boldsymbol{\theta}}_i, \tilde{t}_i) = (S^\alpha(\delta))^i$:

$$\widetilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{\theta}_s(\alpha) \quad (4.9)$$

$$\tilde{t}_i = t_i \quad (4.10)$$

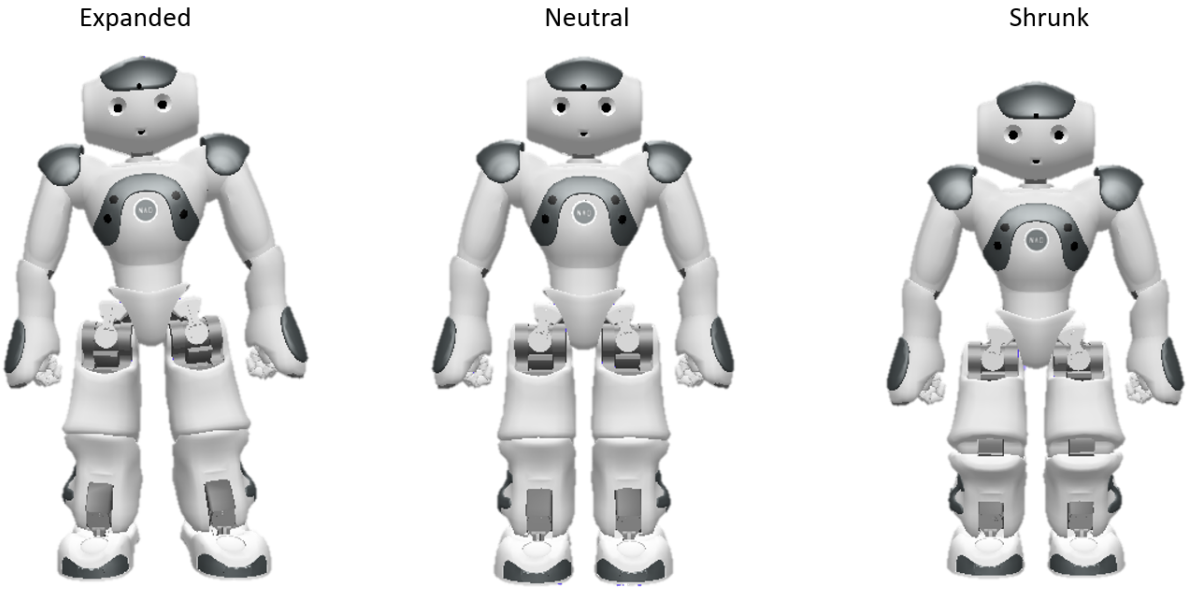


Figure 4.4: An illustration of the the three stances achieved by controlling the stance parameter. The stances portrayed are (left to right) expanded/upright, neutral and shrunk/bend

4.2.1. Order of application

Studies like [16] propose a particular order for applying modifications. By definition, most of the above operators can be applied in any order and generate the same result.

The speed operator modifies the time component of the gesture, whereas amplitude, stance and head pose operators change the poses or angle configurations. Repetition and speed operators are interchangeable as they would generate the same output gesture. Hence, speed can be applied before or after any other operator.

Stance and head pose operators work on mutually exclusive joints and thus are interchangeable. The stance and head pose modifications are applied to every pose in the gesture. Hence, the reference line for these joints would be parallel to the x-axis. Also, all the angle-time points would lie on the reference line and thus would not result in any amplitude change. Since all the poses in a gesture are modified, the repetition operator can also be applied before or after these operators.

In some cases, interchanging amplitude and repetition yields different outputs. Consider an example of a gesture which has two pivot points and some poses after the second pivot point. According to equation 4.1, applying amplitude operator results in 2 reference lines. The amplified output is then repeated by the repetition operator as visualised in figure 4.5(a). On the other hand, applying repetition followed by the amplitude operator results in 4 reference lines, as illustrated in figure 4.5(b). The grey circles highlight the difference in the outputs. Additionally, in this example, applying repetition first produces an asymmetric result because second and fourth reference lines are not parallel. Due to the symmetry, the result of applying amplitude first is deemed desirable. Thus, the only constraint is that the amplitude operator should be applied before repetition.

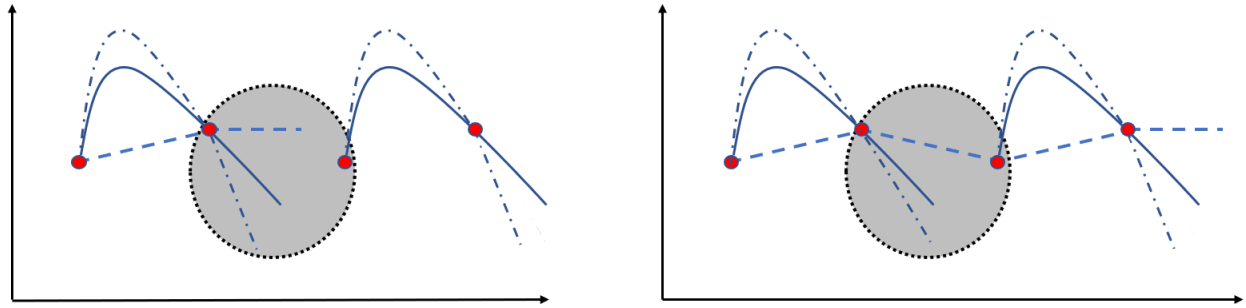


Figure 4.5: An example of amplitude-repetition order resulting in different output. The first graph (a) shows the result of applying amplitude operator followed by repetition. The second graph (b) shows the result of applying repetition followed by amplitude operator.

As argued above, any order which applies amplitude before repetition would produce the same result. Formally, the order followed in the proposed framework is as follows:

$$\delta_{affect} = S^{\alpha_4}(H^{(\alpha_3,\ flag)}(V^{\alpha_2}(R^k(A^{\alpha_1}(\delta))))) \tag{4.11}$$

4.3. LED operator

In addition to the motion and body language operators mentioned in the previous section, this work proposes the use of the LED channel to express affect. Studies like [14, 28] have demonstrated that features like LED colours, blinking frequency and patterns can be used to express various affects in a robot. The definitions of these features are as follows:

1. Colour: Many humanoid robots have LEDs which use a combination of red, green and blue channels to produce a wide variety of colours. This feature uses RGB representation of a colour where each of R, G and B values lie in the range [0, 255].

2. Blink period: The blink period is the time taken to reach the maximum intensity of the intended colour from no colour and back. For fast blinking, the blink period is short, causing rapid switches between no colour and intended colour. Similarly, slow blinking has long blink periods.

3. Rise and fall patterns: [14, 28] demonstrated that time taken to achieve the colour is an important feature for modelling led patterns. Rise-time is the fraction of the blink period taken to reach the maximum intensity, and fall-time is the fraction of the period to reach zero intensity or no colour. Hold-time is the fraction of the period in which the intensity remains maximum. Figure 4.6 illustrates the blink period, rise-time, fall-time and hold-time.

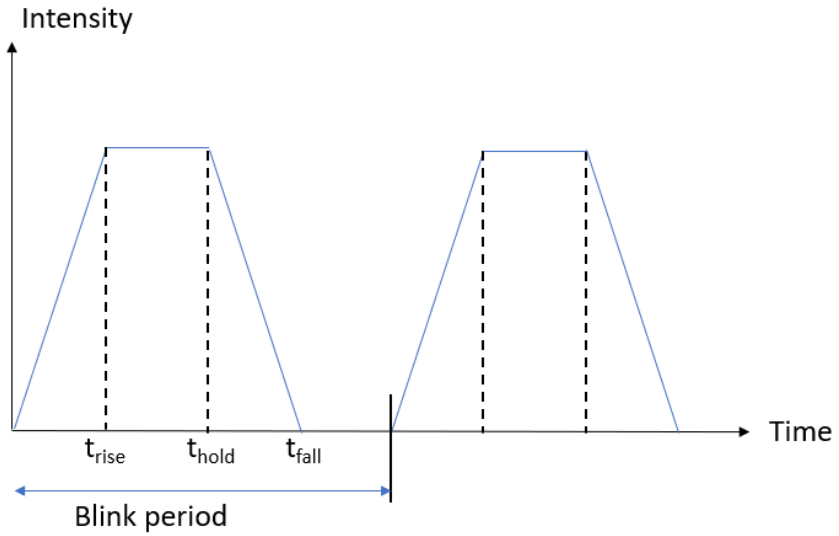


Figure 4.6: A waveform representation of a blink indicating blink period, rise time, fall time and hold time

An LED pattern is a sequence of colour-time pairs that emulates blinking. The key timestamps t_{rise}, t_{hold}, t_{fall}, $T \in \mathbb{R}^+$ ($0 \notin \mathbb{R}^+$) and the associated colours can be interpolated to display an LED pattern. From figure 4.6, it can be observed that between t_{rise} and t_{hold} the intensity is maximum, and between t_{fall} and the blink period T the intensity is zero (or no colour). An LED pattern for on period is written as $\boldsymbol{l} = \langle (RGB,\ t_{rise}),\ (RGB,\ t_{hold}),\ (000,\ t_{fall}),\ (000,\ T) \rangle$, where t_{rise}, t_{hold}, t_{fall} and T follows the constraint $t_{rise} <= t_{hold} <= t_{fall} <= T$.

$colour()$ is a pre-defined function which calculates the RGB value depending on the parameter α.

$$RGB = colour(\alpha)$$

The key timestamps t_{rise}, t_{hold}, t_{fall}, and T belongs to their respective pre-defined ranges $[\tau_r^{min},\ \tau_r^{max}]$, $[\tau_h^{min},\ \tau_h^{max}]$, $[\tau_f^{min},\ \tau_f^{max}]$ and $[\tau_t^{min},\ \tau_t^{max}]$, which follows the constraints $\tau_r^{max} <= \tau_h^{max} <= \tau_f^{max} <= \tau_t^{max}$ and $\tau_r^{min} <= \tau_h^{min} <= \tau_f^{min} <= \tau_t^{min}$. The pre-defined functions $period(), riseRatio(), holdRatio()$ and $fallRatio()$ helps calculate the key timestamps using the parameter β.

$$T = period(\beta) \qquad t_{rise} = riseRatio(\beta) \times T$$
$$t_{hold} = holdRatio(\beta) \times T \qquad t_{fall} = fallRatio(\beta) \times T$$

The LED pattern should be repeated for the duration of the gesture. Similar to gestures, the concatenation of two LED patterns written as $l_1 + l_2$, has the timestamps of the second pattern increased by the final timestamp of the first pattern. Due to similarities in structure and operations, the repetition operator R^k can be re-used for repeating LED patterns where k

denotes the number of times the pattern is repeated. The LED pattern operation $L^{(\alpha,\ \beta,\ duration)}$ yields:

$$k = \left\lfloor \frac{duration}{T} \right\rfloor$$

$$\tilde{\boldsymbol{l}} \quad = \quad R^k(\boldsymbol{l}) \tag{4.12}$$

4.4. Pose repertoire

When using pose repertoires for emotion expression, a pre-defined *key pose* is enacted by the robot to portray a specific emotion. The key pose can be viewed as a singleton gesture $\delta_{key} = \langle(\boldsymbol{\theta^*}_{key},\ \tau)\rangle$, where $\boldsymbol{\theta^*}_{key}$ denotes the key pose associated with the emotion. The time $\tau \in \mathbb{R}^+$ $(0 \notin \mathbb{R}^+)$ is the time by which the robot achieves the pose. Note that $\boldsymbol{\theta^*}_{key}$ is marked as a pivot pose. Applying amplitude operator to a gesture, generated by concatenating key pose with another gesture, could distort the key pose. Marking it as a pivot pose would prevent changes in angles during amplitude modulation.

5

Rendering affect

This chapter describes the parameter values for instantiating the operators introduced in chapter 4. As discussed earlier in chapter 1, this thesis uses a two-dimensional representation of affect. An affect is represented as a point (x, y) in the *valence-arousal* plane, where x and y denote the valence and the arousal values, respectively. The valence and arousal values of all the affects are in the range $[-1, 1]$. There are no fixed values of valence and arousal associated with each affect, i.e. the affects map to areas rather than points. Since the framework requires a point in valence-arousal space as input, paradigmatic points of the affects are used. Russell's circumplex model [24] provides a tentative mapping of 28 affective words. [27] studied affective videos to develop an emotion subspace model as illustrated in figure 5.1. A paradigmatic point was chosen for each affect such that, it is close to the tentative point in [24], and it lies approximately in the centre of the emotion subspace. Table 5.1 lists the eight paradigmatic points which were chosen by comparing the two models.

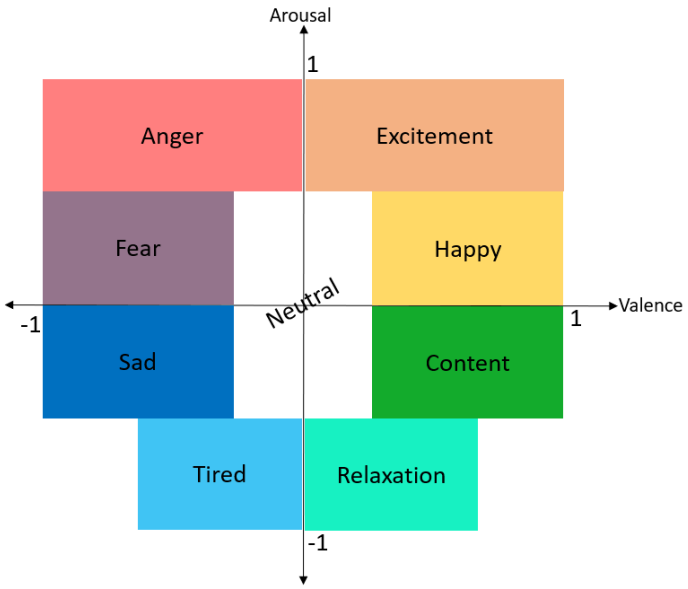


Figure 5.1: An adaptation of emotion sub-spaces in the valence-arousal plane presented in [27]

Studies like [9, 21, 34] have demonstrated that the features can be modulated by valence and arousal values. A parametric affect expression model can be developed by using the values in table 5.1 to determine the parameter values of the various operator discussed in chapter 4. In the following sections, x denotes a valence value in the range [-1, 1] and y

Affect	Valence	Arousal
Neutral	0	0
Happy	0.65	0.3
Excited	0.4	0.75
Anger	-0.2	0.75
Fear	-0.75	0.3
Sad	0.75	-0.35
Tired	-0.3	-0.75
Relaxed	0.3	-0.6
Content	0.7	-0.25

Table 5.1: Paradigmatic valence and arousal values for the affects considered in this thesis

denotes an arousal value in the range [-1, 1]. An affect can be defined in terms of valence and arousal. The idea is to define functions which map valence-arousal values to the parameter values.

5.1. Motion and body language operators

1. Amplitude: [34] found a correlation between the amplitude of motion and the valence of the displayed affect. The amplitude operator discussed in section 4.2 relies mainly on the *pivot points* and amplitude factor α_1. While pivot points are marked in the input gesture, α_1 is passed as a parameter to the amplitude operator. As seen in [9, 21, 30, 34, 35], negative affects have reduced amplitudes ($\alpha_1 < 1$), whereas positive affects have increased amplitudes ($\alpha_1 > 1$). The amplitude factor is clipped to $[0.5,\ 2]$, where $\alpha_1 = 1$ results in no change in amplitude. This range was empirically determined from the values used in [35]. Given the valence value x, the amplitude factor α_1 is computed as:

$$\alpha_1 = \begin{cases} 1 - 0.5x & \text{if } x \leq 0, \\ 1 + x & \text{if } x > 0 \end{cases} \tag{5.1}$$

2. Repetition: As demonstrated in [21, 30], positive arousal is associated with an increase in the repetition of the gesture. On the contrary, negative arousal does not change the repetition of the gesture. The repetition operator takes a positive integer k as the parameter, which can be computed from arousal y as:

$$k = \begin{cases} 1 + \lfloor 2y \rceil & \text{if } y > 0, \\ 1 & \text{otherwise} \end{cases} \tag{5.2}$$

$\lfloor\ \rceil$ denotes rounding to the nearest integer operation. $k = 1$ is the base case where the gesture is performed once. Depending on arousal of the affect, the gesture will be performed 1, 2 or 3 times. This modulation and the repetition values are adopted from [30].

3. Speed: Studies like [9, 21, 30, 34, 35] have demonstrated that speed influences the perceived arousal. An increase in speed portrays high arousal, whereas a reduction in speed portrays low arousal. The velocity operator uses the parameter α_2 is clipped to the range $[0.5,\ 2]$, which was determined empirically from the values used in [35]. The parameter is computed based on arousal y as:

$$\alpha_2 = \begin{cases} 1 - 0.5y & \text{if } y \leq 0, \\ 1 + y & \text{if } y > 0 \end{cases} \tag{5.3}$$

4. Head up-down: The vertical head pose is an important feature for expressing affects in the first quadrant (x > 0, y > 0) and third quadrant (x < 0, y < 0) [3, 13, 33, 34].

The affects in the first quadrant ($Q1$) like happy and excited are associated with head up poses whereas, affects in the third quadrant ($Q3$) like sad and tired are associated with head down poses. This thesis adopts the modulation proposed in [33], which uses valence to determine the vertical head position. The vertical head pose operator expects head pitch factor α_3 and a flag indicating the up or $down$ direction. Given valence x and arousal y, the parameters are computed as:

$$\alpha_3,\ flag = \begin{cases} |x|,\ up & \text{if } (x,\ y) \in Q1, \\ |x|,\ down & \text{if } (x,\ y) \in Q3, \\ 0,\ none & \text{otherwise} \end{cases} \tag{5.4}$$

The pre-defined angles up_{max} and $down_{max}$ for the Nao robot are set to 0.5 and -0.35 radians, respectively. These values were determined empirically from [30].

5. Stance: As noted in [9, 21], the stance of the robot (expand vs shrink) portrays arousal. Stance is classified as an arousal-oriented feature which can be determined solely by the arousal value. Hence, all affects with same arousal use the same stance. The parameter α_4 controls the stance and, given an arousal y, it is computed as:

$$\alpha_4 = y \tag{5.5}$$

The pre-defined leg joint angles used to achieve expanded, neutral and shrunk stances for Nao are adopted from [22] and can be found in table 5.2. The thresholds τ_1 and τ_2 for α_4 are set to 0.5 and -0.5, respectively. So, the robot adopts an expanded stance for arousal values ≥ 0.5 and a shrunk stance for arousal values ≤ -0.5.

Joint name	Expand	Neutral	Shrink
LHipYawPitch	-0.17	0.0	0.0
RHipYawPitch	-0.17	0.0	0.0
LHipRoll	0.09	0.0	0.0
RHipRoll	-0.09	0.0	0.0
LHipPitch	0.13	0.0	-0.44
RHipPitch	0.13	0.0	-0.44
LKneePitch	-0.08	0.0	0.69
RKneePitch	-0.08	0.0	0.69
LAnklePitch	0.08	0.0	-0.35
RAnklePitch	0.08	0.0	-0.35
LAnkleRoll	-0.13	0.0	0.0
RAnkleRoll	0.13	0.0	0.0

Table 5.2: The various leg joints of Nao robot and their corresponding angles to achieve expanded, neutral and shrunk stances. All angles are in radians.

5.2. LED operator

A significant aspect of designing LED patterns is finding suitable colours for expressing affects. Studies like [10, 19] found relationships between colours and emotions. [14, 28] used LED patterns to portray emotions. However, these studies use emotion-specific models, i.e. they used specific colours to express emotions. [10] mapped a hue-circle to the valence-arousal plane, which demonstrates the possibility of developing a parametric model for LEDs. This model has not been tested for affects that are very close in the valence-arousal space. Hence, such affects may not be distinguished by the users through LED patterns. The hues produced by this model for specific emotions, belonging to different parts of the valence-arousal space, resemble the colours found in [14, 19, 28].

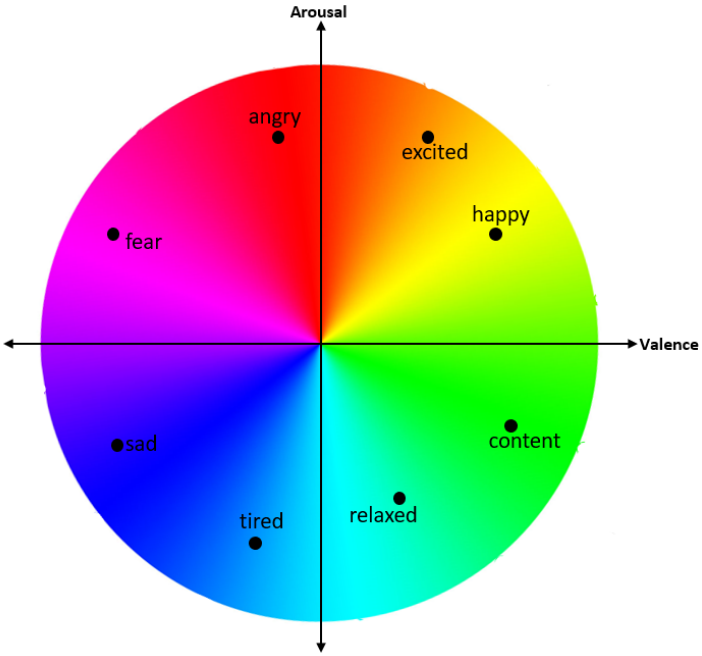


Figure 5.2: Hue circle mapped onto the valence-arousal plane along with the 8 affects from table 5.1.

Colour The affect to colour mapping used in this thesis is inspired by [10]. This mapping is consistent with [19, 28], which associate hues of blue with low arousal, hues of red with high arousal and hues of green with positive valence. Figure 5.2 illustrates a mapping of the hue-circle onto the valence-arousal plane and gives an idea of the hue associated with the affects that are being studied.

The hues on the hue circle corresponds to an angle in the range $[0°, 360°]$, where $0°$ corresponds to red (RGB value $[255, 0, 0]$). So, computing the angle made by the (affect) point with the arousal axis gives the hue corresponding to the affect. The hue can be used to generate the colour in HSV (Hue, Saturation, Value) format. In this case, saturation and value are set to 100%. The HSV format can easily be converted to RGB if required.

The $colour()$ function uses α to generate the RGB value of the colour used in the LED pattern. In this case, the parameter α is the point (x, y) in the valence-arousal plane.

$$\alpha = (x, y) \tag{5.6}$$

The $colour()$ function computes the angle made by the point with the y-axis (arousal axis), which is then used as hue to find the RGB value.

Period As noted in [28], higher arousal is often characterised by fast blinking or short period. The minimum and maximum threshold is set as $\tau_t^{min} = 0.4s$ and $\tau_t^{max} = 4s$, clipping period of any affect to $[0.4, 4]$ seconds. The thresholds were found empirically from the values used in [28]. The $period()$ function uses the parameter β to calculate the blink period. Given an arousal value y, β is calculated as:

$$\beta = \frac{1 - y}{2} \tag{5.7}$$

The period of the LED pattern is computed as:

$$T = period(\beta) = (4 - 0.4) \times \beta + 0.4 \tag{5.8}$$

Rise and fall patterns As discussed in the previous chapter, fractions of the period are reserved for LEDs to reach the maximum intensity and then fall to zero intensity, emulating a blink waveform. The functions $riseRato(), holdRatio(), fallRatio()$ and $period()$ computes the four key timestamps (t_{rise}, t_{hold}, , t_{fall}, T) to generate the intended LED patterns. The computations used in these functions are inspired by [28] and can be defined as follows.

$$riseRatio(\beta) = (0.5 - 0.1) \times \beta + 0.1 \tag{5.9}$$

t_{rise} is the time taken to reach the maximum intensity. Equation 5.9 shows that the maximum of rise-time $\tau_r^{max} = 0.5T$ and the minimum $\tau_r^{min} = 0.1T$. This implies that t_{rise} always lies between 0 and $\frac{T}{2}$, i.e. $0 < t_{rise} \leq \frac{T}{2}$.

$$holdRatio(\beta) = 0.5 \tag{5.10}$$

The hold ratio, i.e. the fraction of period until which the LED stays at maximum intensity, is set to 0.5. Hence, $\tau_h^{min} = \tau_h^{max} = 0.5T$. This implies that in all the cases, the intensity of the LED starts dropping at time $\frac{T}{2}$.

$$fallRatio(\beta) = riseRatio(\beta) + 0.5 \tag{5.11}$$

t_{fall} is the timestamp at which the intensity reaches zero or no colour. This timestamp lies between $\frac{T}{2}$ and T, i.e. $\frac{T}{2} < t_{fall} \leq T$. Equation 5.11 implies that the maximum and minimum achievable fall-time are $\tau_r^{max} = T$ and $\tau_r^{min} = 0.6T$ respectively.

The maximum and minimum values of the key timestamps should adhere to the constraints: $\tau_r^{min} <= \tau_h^{min} <= \tau_f^{min} <= \tau_t^{min}$ and $\tau_r^{max} <= \tau_h^{max} <= \tau_f^{max} <= \tau_t^{max}$. Substituting the corresponding minimum and maximum values, it can be seen that these values conform to the constraint, i.e. $0.1T < 0.5T < 0.6T < T$ and $0.5T = 0.5T < T = T$

A clear difference in LED patterns of various affects is the distinct hue associated with each affect. But, variations in the period, rise and fall ratios also change the blink pattern significantly, as illustrated in figure 5.3. The first pattern resembles a rectangular waveform with a sudden rise and fall in intensities, which is characteristic of high arousal. The second pattern shows a slow rise, long hold and a slow fall, which is associated with low arousal.

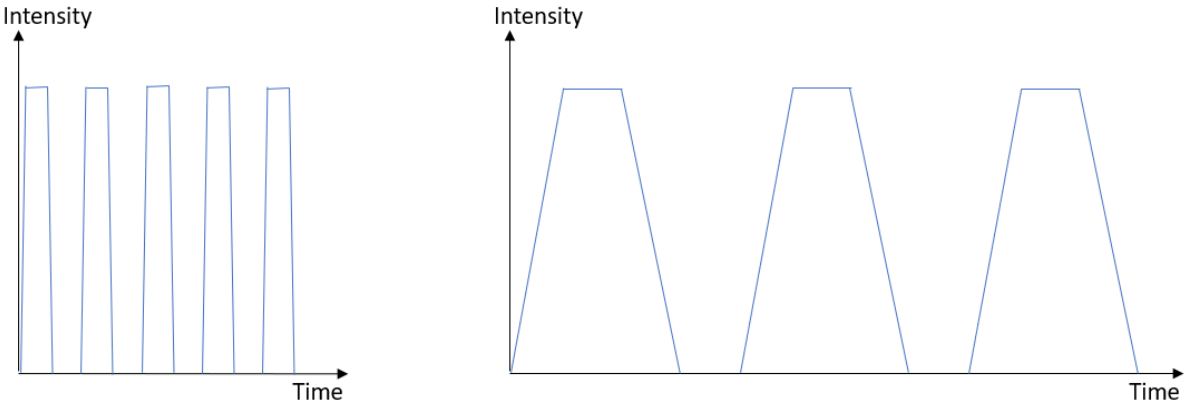


Figure 5.3: Two different blink patterns generated by varying period, rise-time and fall-time

Duration The LED operator uses $duration$ to determine the number of repetitions of the basic LED pattern. Ideally, this parameter equals the duration of the gesture, but it can be set to a longer time if required. In our case, this was set to 8.5 $seconds$ as discussed in chapter 6.

5.3. Pose repertoire

As discussed in chapter 1, *fear* and *content* may not be recognised by using the parametric model based on LED, motion and body language features. While content lacks a well-known key pose, studies like [3, 7] developed key poses which were recognised as fear. Figure 5.4 illustrates the fear pose used in this study.

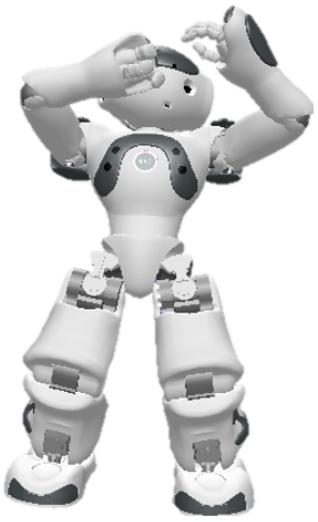


Figure 5.4: The key pose designed for fear, enacted by Nao. The pose was inspired by [3, 7]

6

Experiment Design

6.1. Gestures

Gestures are an integral part of social interactions. People use gestures in their day to day interactions, in group settings like narrations or face-to-face conversations. Hence, it is beneficial to test the affect expression framework on gestures than other tasks like dancing or sports. Many gesture classifications have been proposed throughout the years [15]. This thesis uses the gesture classes proposed in [18] namely, *iconic, metaphoric, beats* and *deictic* gestures.

6.1.1. Iconic

Iconic gestures enact the scenario or physical form of the accompanying speech. For example, enacting size and shape of an object, like drawing a square in the air to portray a square-shaped box. The iconic gestures studied here include:

1. Wave: This a common gesture which accompanies the word *Hello* and is often seen as the enactment of the greeting itself. Figure 6.1 illustrates the poses involved in generating wave gesture.

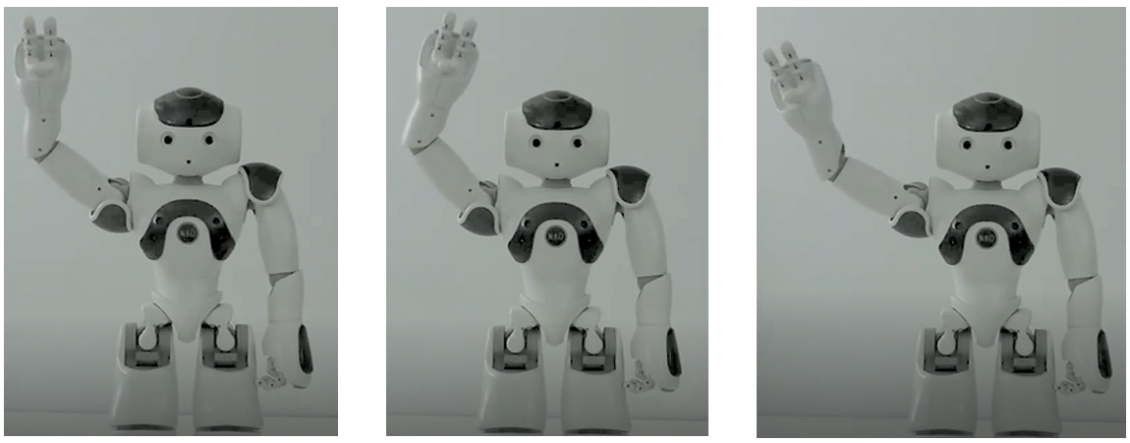


Figure 6.1: The 3 poses which were interpolated and repeated to form the wave gesture

2. Look-around: This gesture is an enactment of looking around by turning the head from left to right. It could be used in scenarios like looking for something or scanning the room. Figure 6.2 illustrates the poses used to generate this gesture.

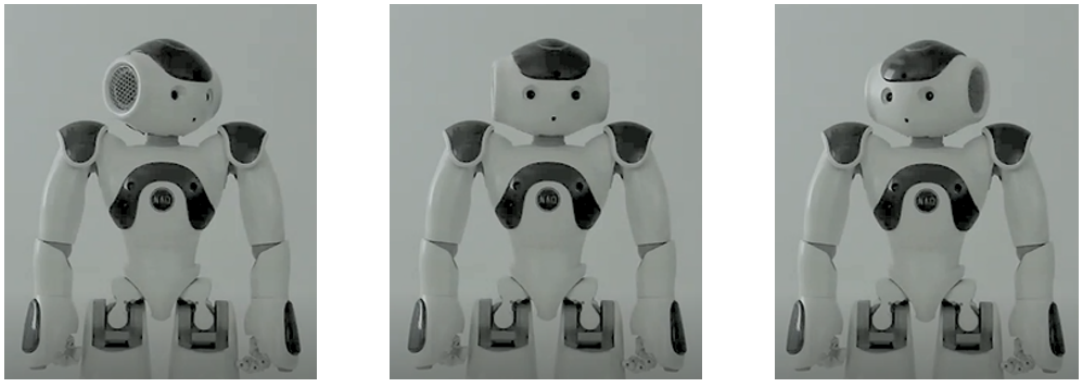


Figure 6.2: The 3 poses which were interpolated to produce look-around gesture

3. Handshake: This gesture is the enactment of shaking someone's hand, which in western culture, is a gesture performed when meeting someone. Figure 6.3 shows the poses involved in a handshake gesture.

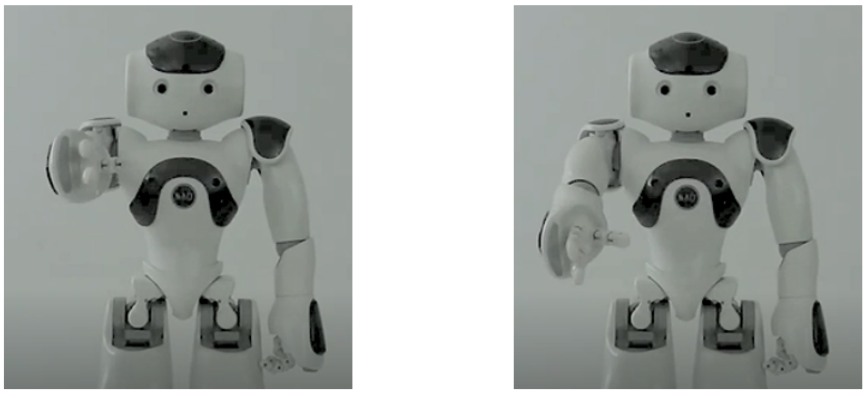


Figure 6.3: The 2 poses which were interpolated and repeated to generate handshake gesture

6.1.2. Metaphoric

Metaphoric gestures represent some abstract concept rather than an enactment of the speech. For example, lifting the index finger and middle finger to form a 'V' often represents victory. Metaphoric gestures can be used without an accompanying speech. The metaphoric gestures used in the experiments are:

Figure 6.4: The 2 poses which were interpolated and repeated to produce nod-yes gesture

1. Nod-yes: This gesture involves moving the head up and down along a vertical line. As the name suggests, this gesture portrays the concept of agreement or validation and is

often used in conversations as a non-verbal gesture. Figure 6.4 illustrates the poses involved in generating nod-yes gesture.

2. Clap: The clap gesture is often used as a form of non-verbal appreciation or lauding. It involves bringing the palms together repeatedly to produce a sound. Figure 6.5 illustrates the hand poses involved in a clapping gesture.

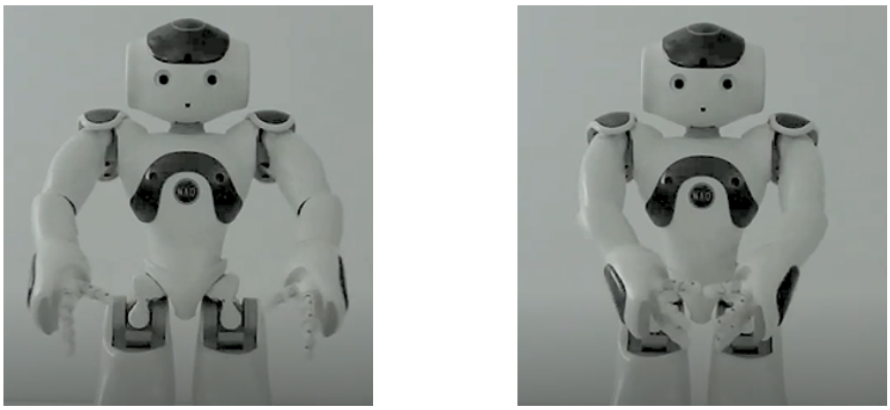


Figure 6.5: The 2 poses which were interpolated and repeated to produce clap gesture

6.1.3. Deictic

Deictic gestures are used to give directions or reference an object. These gestures have a specific purpose, and hence have fewer gestures classified into this category. The deictic gesture considered here is:

1. Pointing: This study uses the pointing forward version of the gesture. It involves lifting the hand, followed by stretching it forward. Figure 6.6 shows the key poses involved in pointing gesture.

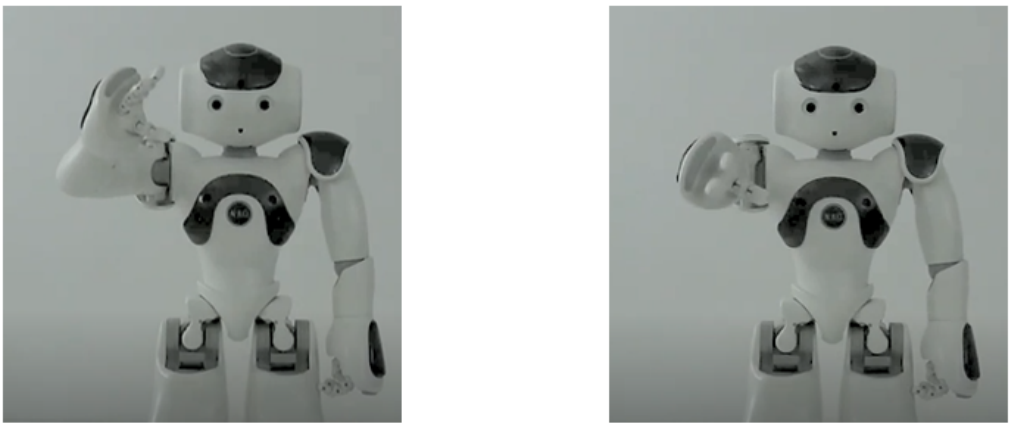


Figure 6.6: The 2 poses which were interpolated to produce pointing gesture

6.1.4. Beats

Beat gestures are simple hand movements used to aid the flow of speech. They convey minimal or no information. Small and repeated hand movements are typical examples of beat gestures. These gestures always co-occur with speech. The following beat gestures are selected for the study.

1. *These* gesture: This gesture can be used while talking about a certain set of objects. The enactment of this gesture involves small vertical movements of half-stretched hands, as illustrated in figure 6.7.

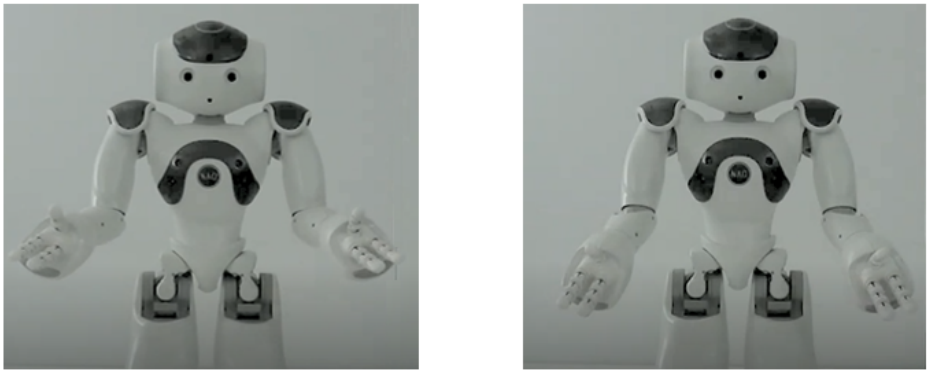


Figure 6.7: The 2 poses which were interpolated and repeated to produce *these* gesture

2. This-or-that gesture: This-or-that gesture involves hand movements used while talking about two objects. Figure 6.8 shows the key poses involved in the gesture.

Figure 6.8: The 3 poses which were interpolated to produce this-or-that gesture

6.2. Experiment setup

The experiment aims to determine which affects are easier to perceive and which model is best suited for expressing each affect. An additional channel is employed only when the affect was not expressible by the previous model. Hence, the experiment was conducted in three phases. The first phase employed the motion and body-language features to express affect. The second phase studied the impact of adding LED features. The third phase expressed emotions through pose repertoires. All phases of the experiment were conducted on *Amazon Mechanical Turk*[1] platform. Table 6.1 lists the models used in each phase along with the affects that were tested. The affects for each phase were determined based on the results of the previous phase. Chapters 7 and 8 discusses the reasons behind the decisions.

6.2.1. Participants

The participants accessed the experiments on Amazon Turk as *HITs*, which awarded financial compensation for successful completion. The participants received textual instructions about the task and what they had to do. Data about *Age, Gender* and *Country* were collected from the participants at the beginning of the experiment. There were no constraints

[1] `https://www.mturk.com/`

Phase	Models	Affects tested
1	Motion and Body language	All - happy, excited, sad, tired, anger, fear, content, relaxed
2	Motion and Body language + LED patterns	anger, fear, content, relaxed
3	Pose repertoire	fear

Table 6.1: The experiment plan for each phase. The second column lists the models used in each phase. The last column lists the affects tested in each phase.

on the age or gender of the participants. However, only participants from North American and European countries could attempt the HIT.

Phase 1 A total of 264 participants (33 participants per gesture × 8 gestures) were recruited through Amazon Mechanical Turk for phase 1. The gender distribution of the participants were: *Male - 111, Female - 151, Other - 1, Prefer not to say - 1.* The participants were aged *18 - 64 years (mean = 36.5 years).* The participants were mostly from the U.S.A *(U.S.A - 245, Canada - 12, U.K - 7).* The demographic details of participants who judged each gesture can be seen in table B.1 .

Phase 2 After phase 1 analysis, 4 affects were considered for phase 2. To keep the load same as phase 1, each batch of participants judged all variations of 2 gestures. For this phase, 132 participants (33 participants per 2 gesture × 8 gestures) were recruited. The gender distribution of the participants were: *Male - 76, Female - 55, Other - 1.* The participants were aged *19 - 68 years (mean = 34.9 years)* and hailed mostly from the U.S.A *(U.S.A - 125, Canada - 6, U.K - 1).* The demographic details of participants who judged each gesture can be seen in table B.2 .

Phase 3 Only one affect was tested in this phase. A small study involving 10 participants was conducted. The participants *(male - 6, female - 4)* belonged to the age range of *21 - 37 years (mean = 26.9 years)* and hailed from the *U.S.A (8)* and *Canada (2).*

Payment and Rejections The estimated load for each participant was approximately 7 minutes. Given the demographics of all Amazon Turk workers, a large number of participants were expected from the U.S.A. The current federal minimum wage in the U.S.A is 7.25 USD, so the financial compensation for the task was calculated as $\frac{7}{60} \times 7.25 \approx 0.85$ USD.

A few participants (around 10%), did not follow the instructions or provided non-serious responses. They were rejected based on the following criteria.

1. Incorrect completion-code: The Amazon Turk platform requires the participants to enter a completion-code as proof for completing a task. The participants who entered incorrect codes were rejected.

2. Rushed submissions: A minimum time (3.5 $minutes$) was calculated for watching the videos and answering the questionnaires. Participants who completed the task faster than this would not have watched the entire video or would not have read the instructions or examples well. Such rushed submissions were rejected. Only two submissions were rejected based on this criterion.

3. Pearson correlation: This criterion was used in [6], which had a similar experimental setup. Pearson correlation between mean ratings of the videos and individual participant ratings was computed, and submissions with coefficient < 0.15 (same threshold as

[6]) were rejected. Almost all the rejections (around 10% of the submissions) were based on this criterion.

6.2.2. Materials

The gestures were presented to the participants as short videos. The videos had a plain white background and a *Nao* robot in the centre. All the videos had the same frame size (970×563). No other objects were visible in the video. Peripherals like power and network cables were not connected, as these components of the robot could distract the participants. All LEDs were turned off for phase 1 and 3. In phase 2, only the eye LEDs were enabled. All the videos were 8.5 *seconds* long and followed the same routine: a few seconds of the idle pose, followed by the gesture, and again few seconds of the idle pose.

The experiment was presented to the participants as a survey. Each page of the survey had a video, followed by a few questions to obtain the perceived affect. The perceived affect was measured through 2 methods: *forced choice* and *valence-arousal ratings*.

The perceived valence and arousal values were measured through *SAM*(Self-Assessment Manikin) [4], which presents different levels of valence and arousal through images. The experiments used the 7-point scale version of SAM to obtain the valence and arousal ratings. The SAM questionnaire presented to the participants is documented in appendix A.

In the *forced-choice* method, participants had to choose the perceived affect from the list: *Neutral, Excited, Happy, Content/Satisfied, Relaxed, Tired, Sad, Fear* and *Angry*. This provided insights about the recognition rate of the affects, and which affects were perceived as another affect.

6.2.3. Procedure

Phase 1 After filling in the demography details, the participants were introduced to the SAM questionnaire along with definitions and examples. The participants were then asked to rate the valence and arousal of a few affective images from OASIS [2]. The participants who correctly rated the images proceeded to the main experiment, and others were screened out.

The participants were divided into batches, and each batch judged all variations of one gesture. First, the participants watched the video of neutral gesture, i.e. the original gesture without any feature modulations. Next, they viewed the videos of the modulated gestures. Each participant viewed the affective videos in random order. After watching each video, the participants rated the valence and arousal of the robot through the SAM questionnaire. Then, they had to choose the perceived affect through the forced-choice method. This routine was followed for 9 videos (1 neutral + 8 affective) per gesture and all 8 gestures, resulting in a total of 72 variations.

Phase 2 This phase of the experiment was similar to the first phase. The participants filled in the demography details and proceeded to the SAM questionnaire test. Since this phase involved LED colour patterns, the participants had to answer an additional question to ensure that they were not (partially) colour blind.

All aspects of the videos such as length, frame size and background were identical, except the eye LED patterns. Unlike the first phase where all LEDs were turned off, this phase had eye LEDs displaying colour patterns. The LED patterns were repeated throughout the video, even during the idle pose, to avoid too much focus on the patterns.

Some affects were well perceived using motion and body language features. Hence, this phase studied the affects, which were neither recognised well nor rated close to the intended

[2] http://benedekkurdi.com/oasis.php

valence-arousal values. As seen in chapter 7, anger, fear, content and relaxed were considered in phase 2. The videos were distributed among batches such that the participants had task loads similar to phase 1.

Phase 3 As discussed in previous chapters, this work employs emotion-specific pose repertoires as a last resort. Hence, this phase of the experiment was conducted after analysing the previous data. As expected, this phase had very few affects (fear and content). Since we could not find well-known pose for content, only fear was tested in this phase.

This phase evaluated the ease of emotion recognition in key poses. Hence, the measures used were different than other phases. For example, the valence-arousal ratings do not seem suitable for the task. The participants responded to a questionnaire which tested ease of emotion recognition. First, the participants had to describe what they saw in the video. Next, they had to name the emotion that they would associate with the pose. Finally, they had to select the perceived emotion from the provided list (forced-choice). If the initial two responses contained words like afraid, fear, etc. then the pose was deemed easily recognisable.

7

Results - phase 1

The study was conducted in 3 phases, each phase investigating the improvement obtained by employing an additional affect expression technique. This chapter presents and discusses the results obtained in the first phase (detailed in chapter 6).

7.1. Phase 1

The experiments were conducted in batches where each batch consisted of 33 participants. Each batch viewed and judged all 9 versions (1 neutral version and 8 affective versions) of a particular gesture. A total of 264 participants (33 participants per gesture × 8 gestures) were recruited through Amazon Mechanical Turk[1]. For each of the 9 videos, the participants had to rate the valence, arousal and choose an affect which they thought the robot expressed. Since the framework was designed to work with valence and arousal values in the range $[-1, 1]$, the ratings were scaled to the same range for ease of comparison. The following subsections discuss the demographics and results of each batch. The data regarding the demographics were self-reported by the participants.

7.1.1. Wave gesture

Table 7.1 shows the number of participants who recognised the expressed affect along with the mean and standard deviation of the valence-arousal ratings of each video. The full data of affect labels chosen by the participants can be found in table C.1. As hypothesised, *excited* and *sad* have high recognition rates (> 75%) and tired has a medium recognition rate (50-75%). On the other hand, the recognition rate of *happy* is lower than expected.

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	25	0	0.081 ± 0.145	0	-0.061 ± 0.211
Happy	17	0.65	0.383 ± 0.29	0.3	0.192 ± 0.22
Excited	29	0.4	0.374 ± 0.32	0.75	0.636 ± 0.193
Anger	3	-0.2	0.121 ± 0.342	0.75	0.535 ± 0.263
Fear	0	-0.75	0.071 ± 0.298	0.3	-0.192 ± 0.3
Sad	28	-0.75	-0.657 ± 0.27	-0.35	-0.414 ± 0.277
Tired	19	-0.3	-0.343 ± 0.328	-0.75	-0.667 ± 0.186
Relaxed	9	0.3	0.051 ± 0.302	-0.6	-0.505 ± 0.313
Content	6	0.7	0.01 ± 0.195	-0.25	-0.111 ± 0.35

Table 7.1: The table shows the number of participants who recognised the intended affect in each version of the wave gesture. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50% – 75%. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

[1] https://www.mturk.com

Figure 7.1 shows that the observed means of *neutral, excited, sad* and *tired* are close to the intended values. Similar to the recognition rate, the observed mean of *happy* is slightly deviated from the intended value. As seen from figure 7.1 and table C.1, *angry* version of the gesture was mostly perceived as excited or happy, which is in line with the observations of [32]. Interestingly, the arousal ratings of *relaxed* is close to the intended value. A plausible reason for this observation is that, as can be seen in table C.1, *relaxed* was often perceived as *tired* which has a similarly low arousal value. Both *content* and *fear* are farthest from their intended values and have low recognition rates.

Sad and *tired* were rated close to their intended values and belong to the same quadrant in valence-arousal space. A paired samples t-test was done on the arousal and valence ratings of these two affects. There was a significant difference in the scores of both valence ($t(32) = -4.159, p = 0.00022$) and arousal ($t(32) = 4.218, p = 0.00018$). This implies that *sad* and *tired* are distinguishable. Similarly, *excited* and *happy* fall into the same quadrant. Though *happy* is slightly deviated, it is still interesting to check whether they are distinguishable. A paired samples t-test was done on the arousal and valence ratings and a significant difference was found in arousal ($t(32) = 11.065, p = 0.00000$). Hence, these affects are distinguishable based on arousal.

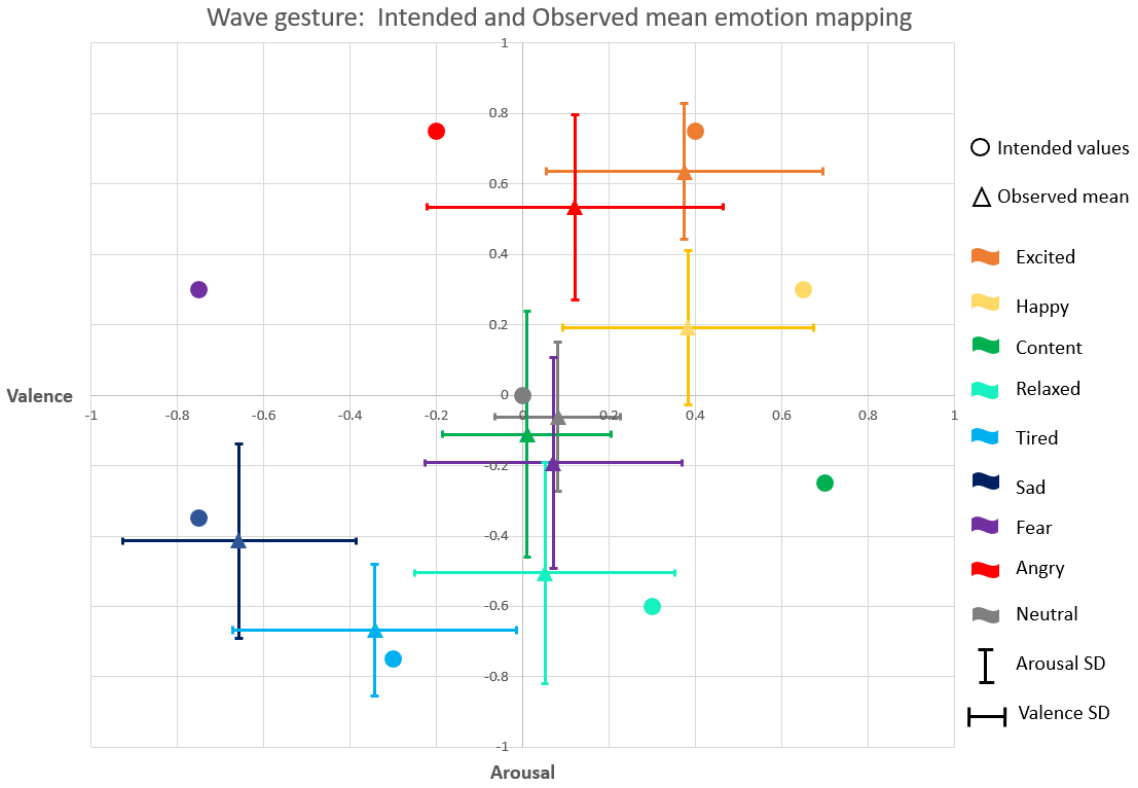


Figure 7.1: Intended values and observed means of affects for the wave gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

7.1.2. Look-around gesture

As seen in table 7.2, only *sad* has a high recognition rate. The recognition rates of *happy* and *excited* is lower than expected, with *happy*'s rate falling under 50%. Interestingly, there is an increase in the recognition of negative affects like *anger* and *fear*. Table C.2 shows that a considerable number of participants perceived *happy* and *excited* as *anger* or *fear*. This is also reflected in figure 7.2 with an over-arching shift towards left in the mean valence ratings, the most shifted affects being *happy* and *excited*. The gesture involves averted gaze or looking away for the majority of its duration, which may cause such a shift. [1, 13, 14] associates

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	23	0	-0.02 $\pm$ 0.203	0	-0.131 $\pm$ 0.263
Happy	9	0.65	0.061 $\pm$ 0.404	0.3	0.303 $\pm$ 0.337
Excited	17	0.4	0.071 $\pm$ 0.389	0.75	0.616 $\pm$ 0.237
Anger	12	-0.2	-0.263 $\pm$ 0.389	0.75	0.596 $\pm$ 0.247
Fear	5	-0.75	-0.111 $\pm$ 0.462	0.3	0.03 $\pm$ 0.394
Sad	27	-0.75	-0.646 $\pm$ 0.372	-0.35	-0.374 $\pm$ 0.273
Tired	13	-0.3	-0.283 $\pm$ 0.302	-0.75	-0.606 $\pm$ 0.328
Relaxed	7	0.3	-0.03 $\pm$ 0.367	-0.6	-0.465 $\pm$ 0.343
Content	3	0.7	-0.01 $\pm$ 0.243	-0.25	-0.222 $\pm$ 0.34

Table 7.2: The table shows the number of participants who recognised the intended affect in each version of the look-around gesture. The cells are coloured green for recognition rates $\geq 75\%$ and cyan for rates between $50\% - 75\%$. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

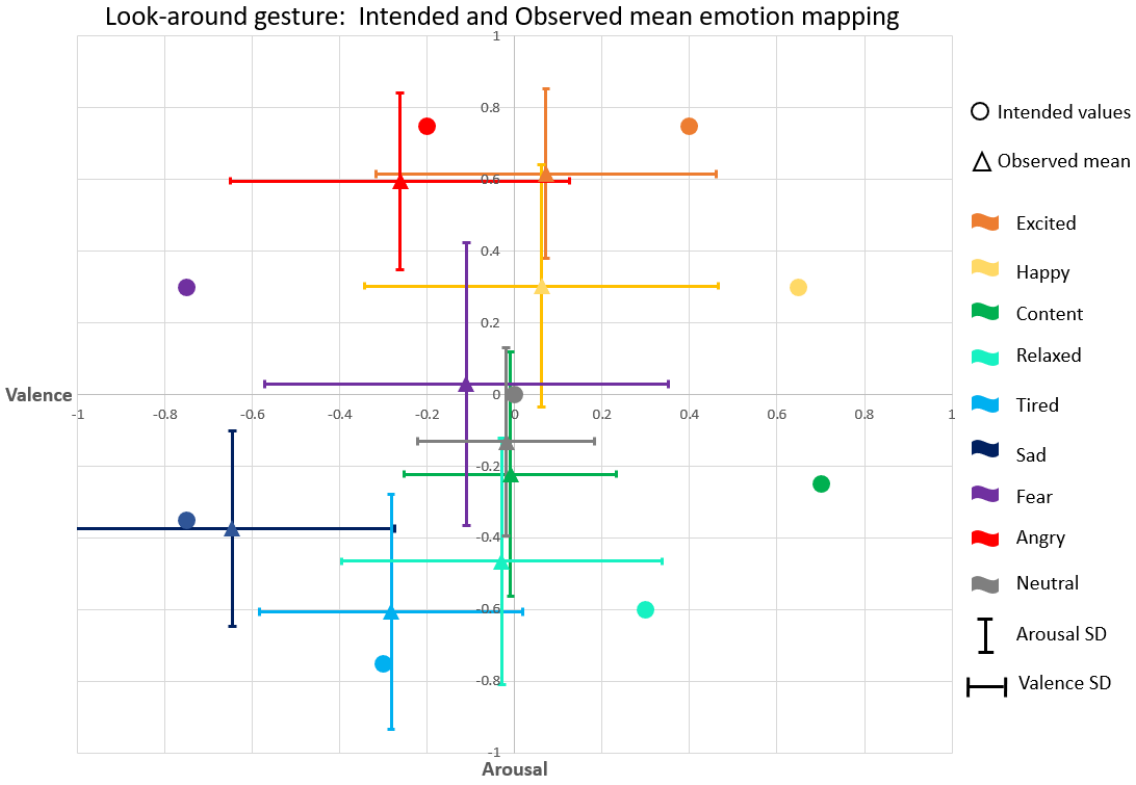


Figure 7.2: Intended values and observed means of affects for the look-around gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

an averted gaze with avoidance-oriented emotions like sad, disgust and fear. These emotions have a negative valence which may have reflected in the overall valence ratings of this gesture.

A paired samples t-test was done on the arousal and valence ratings of *sad* and *tired*. There is a significant difference in both valence ($t(32) = -4.156, p = 0.00023$) and arousal ($t(32) = 3.055, p = 0.0045$). Hence, *sad* and *tired* are distinguishable. The same test was done among *anger*, excited and *happy*. There is a significant difference in arousal ($t(32) = 4.717, p = 0.00004$) of *happy* and *excited* and hence, they can be distinguished along arousal. Similarly, *excited* and *angry* can be distinguished by valence ($t(32) = 5.272, p = 0.00000$). *Angry* and *happy* are distinguishable by both valence ($t(32) = 3.2, p = 0.0031$) and arousal ($t(32) = -4.543, p = 0.00001$).

7.1.3. Handshake gesture

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	26	0	0.101 ± 0.176	0	-0.121 ± 0.183
Happy	11	0.65	0.364 ± 0.226	0.3	0.202 ± 0.288
Excited	24	0.4	0.253 ± 0.334	0.75	0.606 ± 0.212
Anger	2	-0.2	0.263 ± 0.286	0.75	0.374 ± 0.182
Fear	0	-0.75	0.04 ± 0.247	0.3	-0.354 ± 0.343
Sad	26	-0.75	-0.626 ± 0.26	-0.35	-0.455 ± 0.183
Tired	17	-0.3	-0.253 ± 0.323	-0.75	-0.636 ± 0.327
Relaxed	4	0.3	-0.061 ± 0.294	-0.6	-0.434 ± 0.317
Content	4	0.7	0.03 ± 0.226	-0.25	-0.232 ± 0.348

Table 7.3: The table shows the number of participants who recognised the intended affect in each version of the handshake gesture. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50% – 75%. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

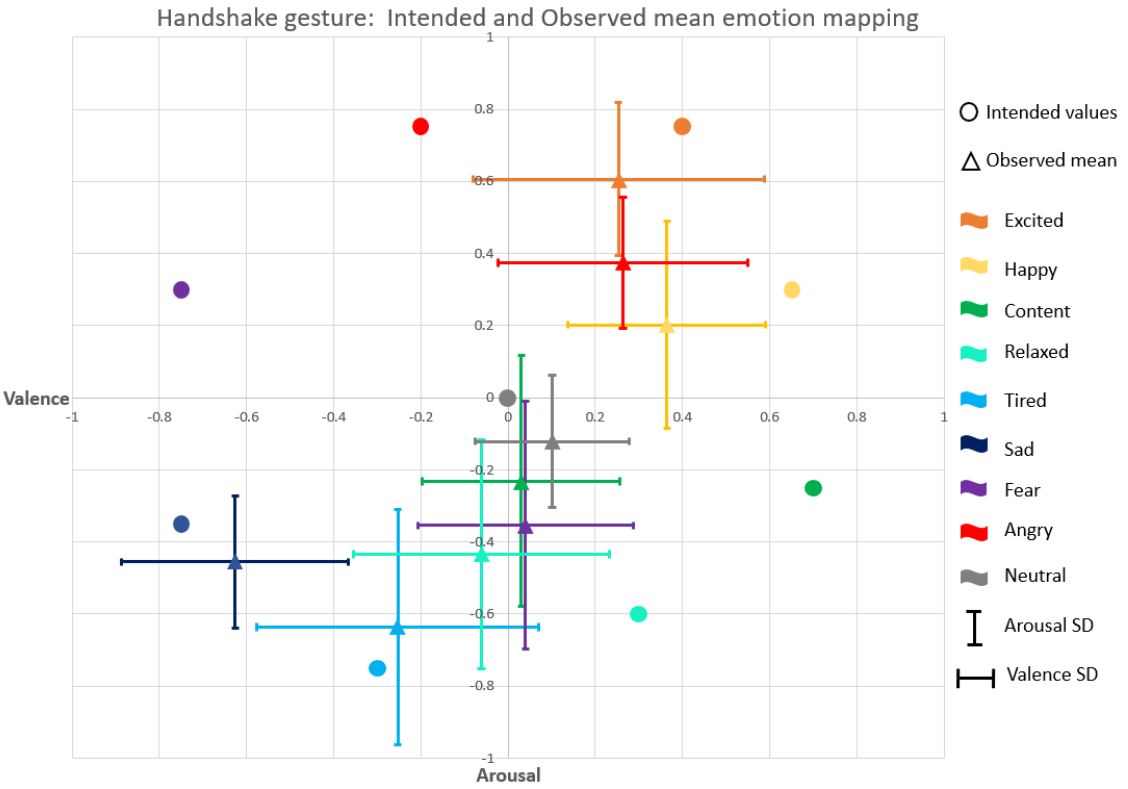


Figure 7.3: Intended values and observed means of affects for the handshake gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

Table 7.3 shows trends similar to the *Wave* gesture. The recognition rates of various affects, except *happy*, conform to the hypothesis. Figure 7.3 shows that the observed means of *excited, sad* and *tired* are close to their intended values. Thus, the ratings of *sad, tired, excited* and *happy* fall into the expected quadrants. The next step is to check if they are distinguishable through a paired samples t-test. *Sad* and *tired* are significantly different along both valence ($t(32) = -5.949, p = 0.00000$) and arousal ($t(32) = 2.667, p = 0.01191$). *Excited* and *happy* are distinguishable by the arousal ratings ($t(32) = 7.25, p = 0.00000$).

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	15	0	0.09 ± 0.172	0	-0.03 ± 0.281
Happy	15	0.65	0.475 ± 0.289	0.3	0.354 ± 0.288
Excited	28	0.4	0.455 ± 0.352	0.75	0.636 ± 0.241
Anger	0	-0.2	0.414 ± 0.289	0.75	0.525 ± 0.264
Fear	0	-0.75	0.101 ± 0.228	0.3	-0.242 ± 0.393
Sad	26	-0.75	-0.374 ± 0.273	-0.35	-0.364 ± 0.241
Tired	14	-0.3	-0.273 ± 0.269	-0.75	-0.596 ± 0.26
Relaxed	7	0.3	-0.101 ± 0.46	-0.6	-0.293 ± 0.398
Content	6	0.7	0.02 ± 0.22	-0.25	-0.232 ± 0.404

Table 7.4: The table shows the number of participants who recognised the intended affect in each version of the nod-yes gesture. The cells are coloured green for recognition rates ≥ 75%. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

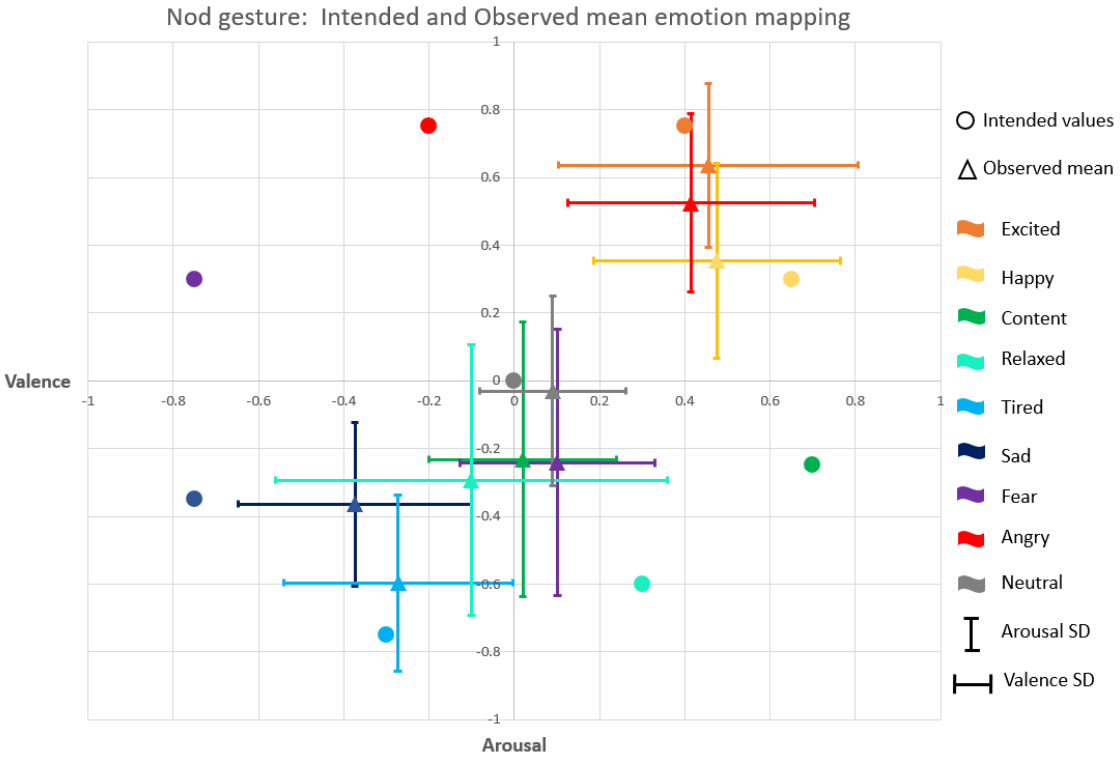


Figure 7.4: Intended values and observed means of affects for the nod-yes gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

7.1.4. Nod-yes gesture

Table 7.4 shows that *excited* and *sad* have high recognition rates. The nod-yes gesture involves the head pitch joint which is also an important body-language feature that conveys valence. In such cases, the model does not alter the joint and thus, hampers its expressiveness. Hence, the affects which are theoretically close like *happy-excited*, *excited-angry*, *sad-tired*, etc. would be very difficult to differentiate along valence. This intuition is proven by figure 7.4, which shows the observed means of various affects cluttered together. *Sad*, a huge beneficiary of head pitch feature, falls considerably far from the intended point on valence-arousal space. However, the affect pairs *sad-tired* and *excited-happy* look separated, making them the best candidates for testing distinguishable affects. A paired samples t-test was run on the valence and arousal ratings of the two affect pairs. A significant difference was found only along arousal for *sad-tired* ($t(32) = 3.538, p = 0.00126$) and *excited-happy* (

$t(32) = 4.855, p = 0.00003$).

7.1.5. Clap gesture

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	18	0	0.172 ± 0.206	0	0.05 ± 0.206
Happy	22	0.65	0.535 ± 0.249	0.3	0.263 ± 0.361
Excited	25	0.4	0.414 ± 0.471	0.75	0.717 ± 0.169
Anger	5	-0.2	0.192 ± 0.312	0.75	0.657 ± 0.228
Fear	2	-0.75	-0.01 ± 0.328	0.3	-0.081 ± 0.334
Sad	25	-0.75	-0.616 ± 0.265	-0.35	-0.444 ± 0.297
Tired	17	-0.3	-0.273 ± 0.269	-0.75	-0.596 ± 0.26
Relaxed	10	0.3	-0.091 ± 0.254	-0.6	-0.545 ± 0.274
Content	9	0.7	0.02 ± 0.249	-0.25	-0.273 ± 0.228

Table 7.5: The table shows the number of participants who recognised the intended affect in each version of the clap gesture. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50% – 75%. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

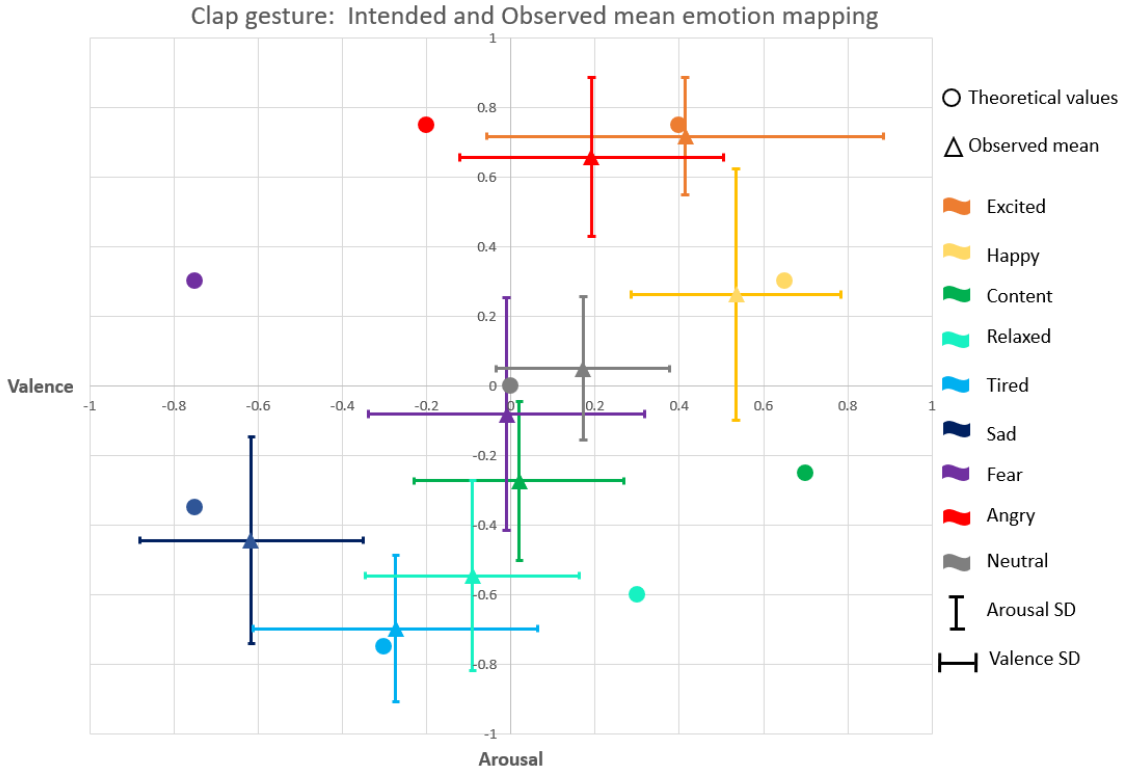


Figure 7.5: Intended values and observed means of affects for the clap gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

As seen in table 7.5,*happy, excited, sad* and *tired* have high recognition rates. Figure 7.5 shows that these affects have mean valence and arousal ratings close the intended values. Interestingly, the observed mean ratings of *happy* were close to intended. As seen in table C.5, the neutral version of the gesture was often recognised as happy. This is reflected in figure 7.5 as a slight shift in mean ratings of neutral towards positive valence. Many people associate clapping with appreciation, which is often perceived as positive. This could have contributed to the higher valence ratings of *happy*.

A paired samples t-test was run on the affect pairs *excited-happy* and *sad-tired*. *Sad* and *tired* are distinguishable along valence ($t(32) = -4.6, p = 0.00006$) and arousal ($t(32) = 4.49, p = 0.00008$). *Excited* and *happy* are significantly different only along arousal ($t(32) = 7.878, p = 0.00000$).

7.1.6. Pointing gesture

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	22	0	0.03 ± 0.153	0	-0.091 ± 0.267
Happy	9	0.65	0.242 ± 0.366	0.3	0.333 ± 0.289
Excited	18	0.4	0.091 ± 0.356	0.75	0.626 ± 0.232
Anger	11	-0.2	-0.071 ± 0.351	0.75	0.657 ± 0.228
Fear	2	-0.75	0.07 ± 0.32	0.3	-0.172 ± 0.383
Sad	22	-0.75	-0.606 ± 0.195	-0.35	-0.394 ± 0.282
Tired	14	-0.3	-0.374 ± 0.232	-0.75	-0.657 ± 0.257
Relaxed	3	0.3	-0.03 ± 0.268	-0.6	-0.475 ± 0.354
Content	3	0.7	0.051 ± 0.252	-0.25	-0.283 ± 0.29

Table 7.6: The table shows the number of participants who recognised the intended affect in each version of the pointing gesture. The cells are coloured cyan for recognition rates between $50\% - 75\%$. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

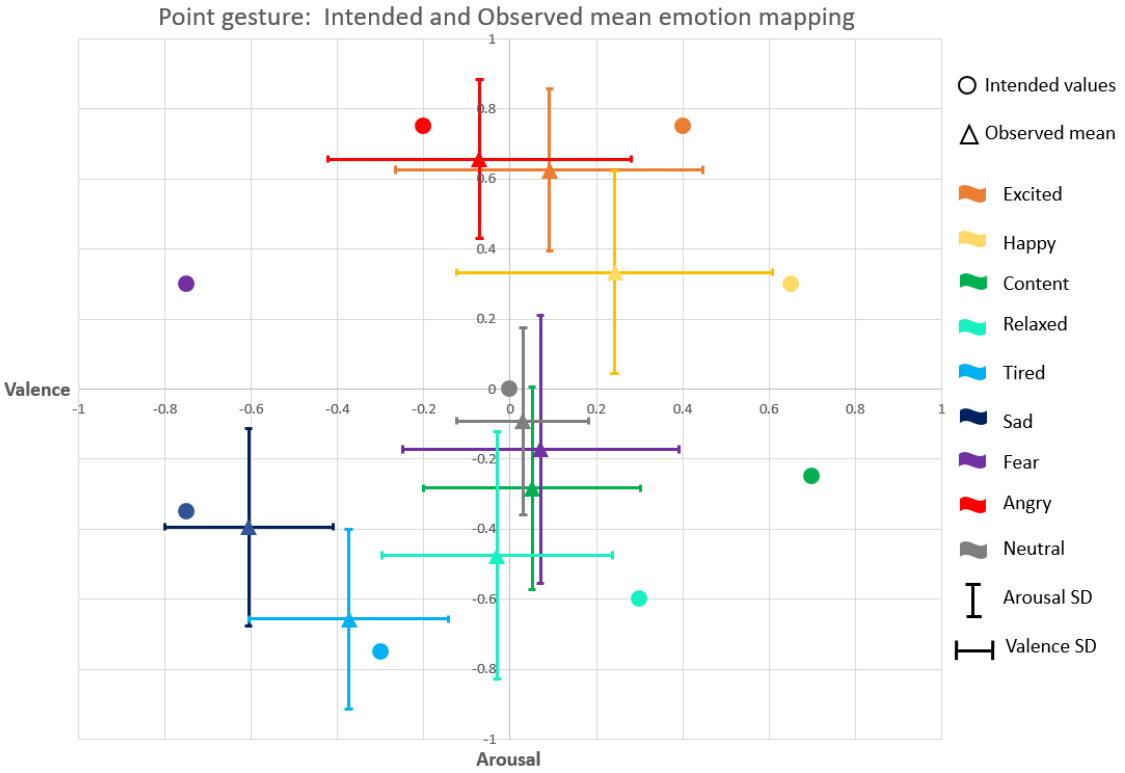


Figure 7.6: Intended values and observed means of affects for the pointing gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

It can be seen from table 7.6 that, unlike other gestures, none of the affects have high recognition rates. However, there is a notable increase in recognition rate of *anger*. Table C.6 also shows a considerable increase in the number of participants who perceived *excited* and *happy* versions as *angry*. The consequence of this is seen in figure 7.6, where *angry* is

closer to the intended point. *Excited* and *happy* have shifted significantly to the left.

A paired samples t-test was run on affect pairs *sad-tired, excited-happy* and *angry-excited*. *Sad* and *tired* showed a significant difference along both valence ($t(32) = -4.208, p = 0.00019$) and arousal ($t(32) = 4.073, p = 0.00029$). *Excited* and *happy* also showed a significant difference along both valence ($t(32) = -2.33, p = 0.02626$) and arousal ($t(32) = 4.662, p = 0.00005$). Also, *anger* is distinguishable from *excited* along valence ($t(32) = 2.369, p = 0.02405$).

7.1.7. *These* gesture

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	23	0	-0.061 ± 0.195	0	-0.081 ± 0.264
Happy	10	0.65	0.283 ± 0.278	0.3	0.232 ± 0.306
Excited	26	0.4	0.262 ± 0.309	0.75	0.596 ± 0.2
Anger	8	-0.2	0.141 ± 0.373	0.75	0.465 ± 0.22
Fear	0	-0.75	-0.061 ± 0.194	0.3	-0.152 ± 0.426
Sad	30	-0.75	-0.727 ± 0.228	-0.35	-0.384 ± 0.278
Tired	10	-0.3	-0.444 ± 0.231	-0.75	-0.475 ± 0.334
Relaxed	8	0.3	-0.03 ± 0.327	-0.6	-0.444 ± 0.297
Content	6	0.7	-0.051 ± 0.302	-0.25	-0.232 ± 0.404

Table 7.7: The table shows the number of participants who recognised the intended affect in each version of the *these* gesture. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50% – 75%. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

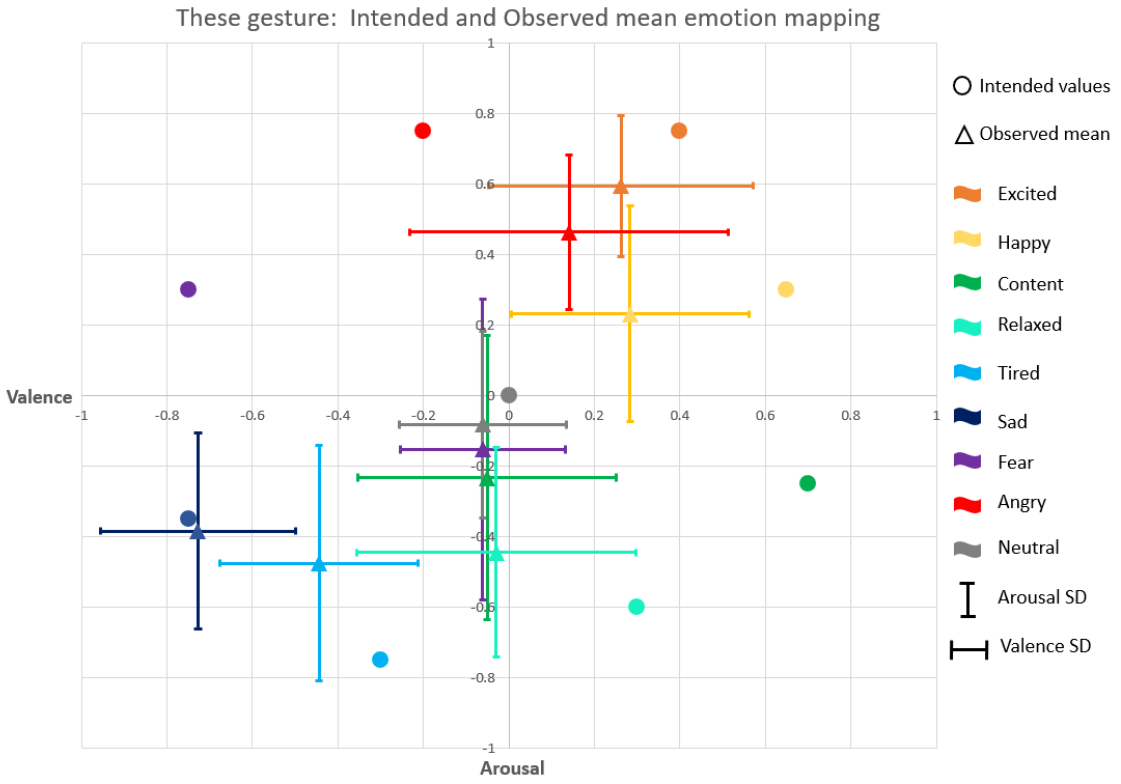


Figure 7.7: Intended values and observed means of affects for the *these* gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

The data of this gesture shows similar trends as other gestures, where *sad* and *excited* have high recognition rates as seen in table 7.7. Figure 7.7 shows that the affects *excited, happy, sad* and *tired* are in the expected quadrant of valence-arousal space. The affect pairs *excited-happy* and *sad-tired* were further analyzed through paired samples t-test to check if they are distinguishable. While *sad* and *tired* are distinguishable along valence ($t(32) = -6.456, p = 0.00000$), *excited* and *happy* are distinguishable along arousal ($t(32) = 5.697, p = 0.00000$).

7.1.8. This-or-that gesture

Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Neutral	22	0	-0.051 ± 0.147	0	-0.101 ± 0.243
Happy	17	0.65	0.293 ± 0.331	0.3	0.242 ± 0.304
Excited	25	0.4	0.232 ± 0.348	0.75	0.576 ± 0.254
Anger	5	-0.2	0.141 ± 0.301	0.75	0.505 ± 0.252
Fear	1	-0.75	-0.131 ± 0.263	0.3	-0.283 ± 0.392
Sad	29	-0.75	-0.687 ± 0.311	-0.35	-0.444 ± 0.198
Tired	13	-0.3	-0.414 ± 0.25	-0.75	-0.677 ± 0.243
Relaxed	2	0.3	-0.071 ± 0.273	-0.6	-0.515 ± 0.313
Content	4	0.7	-0.091 ± 0.209	-0.25	-0.343 ± 0.306

Table 7.8: The table shows the number of participants who recognised the intended affect in each version of the this-or-that gesture. The cells are coloured green for recognition rates ≥ 75% and cyan for rates between 50% – 75%. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

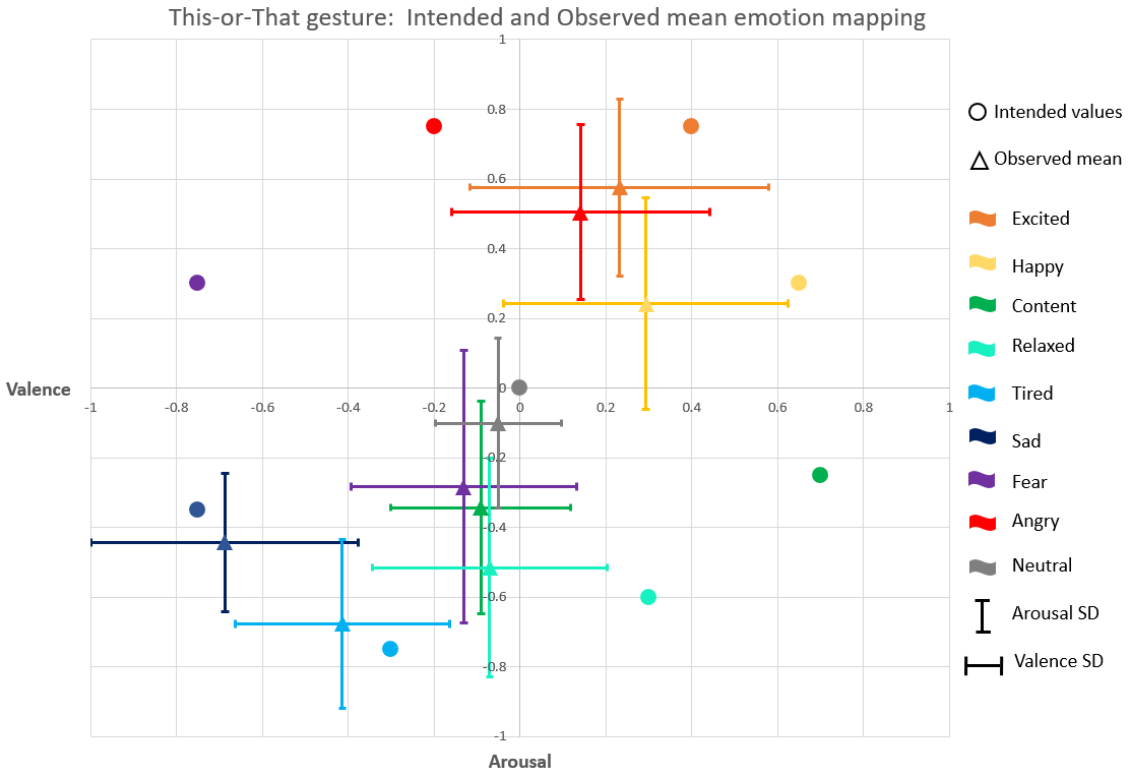


Figure 7.8: Intended values and observed means of affects for the this-or-that gesture, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

The recognition rates are on par with expectations. Figure 7.8 shows that *excited, happy, sad* and *tired* are rated in the expected quadrant of valence-arousal space. A paired samples

t-test was run on these affects. Similar to previous gestures, *happy* and *excited* are distinguishable along arousal ($t(32) = 4.343, p = 0.00013$) whereas *sad* and *tired* are distinguishable along both valence ($t(32) = -4.5, p = 0.00008$) and arousal ($t(32) = 4.946, p = 0.00002$).

7.2. Key results

The above section presented the results and analysis for each gesture. This section highlights the overall takeaways from the results of phase 1 of the experiments. The following trends were observed over various gestures through parametric modulation of motion and body language features.

1. The mean valence and arousal ratings of *excited, happy, sad* and *tired* are in the same quadrant as their intended values.
2. The mean ratings of *excited, sad* and *tired* are close to their intended values in most cases (at least 6 out of 8 gestures). However, the mean ratings of *happy* show slightly lower valence than intended.
3. In all gestures, the affect pairs *excited-happy* and *sad-tired* are distinguishable along at least one of the valence-arousal axes.
4. *Anger* is often recognised as *excited* or *happy* and the valence-arousal ratings reflect this too. Similarly, *relaxed* is sometimes recognised as *tired* which shifts the mean ratings towards lower valence.
5. The mean ratings of *content* and *fear* are the farthest from their intended values. Both these affects have mean ratings close to *neutral*. While *content* received ratings close to *neutral*, *fear* was recognised and rated as different affects by different participants.

8

Results - phase 2 & 3

The results of phase 1 influenced phase 2 of the experiments. The valence-arousal ratings of some affects like *excited, sad* and *tired* were close to the intended values, which indicates that they are easily perceived through the motion and body language modulations. Though the mean ratings of *happy* fall into the intended quadrant, the mean valence was slightly lower than expected. However, studies like [27, 32] consider it as an acceptable valence value of *happy*. It is plausible to obtain mean valence ratings closer to intended values by fine-tuning the valence-oriented operators. For example, studies like [3, 13, 30] demonstrated head-up pose as a strong indicator of happiness. The range of vertical head pose was determined empirically through illustrations in [3, 34]. Hence, this feature could be fine-tuned by using feedback from people in the design phase. The affects: *anger, fear, relaxed* and *content* are considered in phase 2.

8.1. Phase 2

In addition to the motion and body language features, phase 2 employs the LEDs on the robot. Studies like [14, 28] have demonstrated that the colour *red* portrays *anger*. [10, 19, 28] suggests that shades of green portray *relaxed* and *content*, whereas certain hues of purple portray *fear*. Since *anger* and *relaxed* were moderately close to the intended points in phase 1, the ratings of these affects are expected to shift considerably closer to the intended points. On the other hand, though *fear* and *content* may improve in recognition and subsequent ratings, they may still be far from the intended values.

Since this phase studies only four affects, each participant had to evaluate 2 gestures. Hence, each participant judged 8 videos (2 gestures × 4 affects per gesture), thus maintaining the same task load as phase 1. The following subsections present the results for affect pairs which were judged by the same participants.

8.1.1. Clap and Look-around gesture

Table 8.1 shows that the recognition rate of *anger* in both clap and look-around gestures have improved drastically. The recognition rate of *fear* has also improved in both gestures. These observations are also reflected in figure 8.1, which clearly shows the observed mean of anger close to the intended point in valence-arousal space. *Fear* and *content* are rated into the intended quadrants but are moderately faraway from the intended points. The addition of LED patterns for *relaxed* shifted the observed means close to the intended values. As noted in section 7.1.2, the look-around gesture seems to express slightly negative affect plausibly due to the averted gaze embedded into the gesture. This still holds since figure 8.1 shows a small shift towards negative valence for all 4 affects.

Gesture	Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Clap	Anger	18	-0.2	-0.232 $\pm$ 0.348	0.75	0.657 $\pm$ 0.243
	Content	8	0.7	0.303 $\pm$ 0.305	-0.25	-0.091 $\pm$ 0.336
	Fear	7	-0.75	-0.02 $\pm$ 0.311	0.3	0.313 $\pm$ 0.322
	Relaxed	5	0.3	0.293 $\pm$ 0.232	-0.6	-0.343 $\pm$ 0.317
Look-around	Anger	19	-0.2	-0.444 $\pm$ 0.245	0.75	0.434 $\pm$ 0.282
	Content	4	0.7	0.01 $\pm$ 0.328	-0.25	-0.293 $\pm$ 0.273
	Fear	8	-0.75	-0.303 $\pm$ 0.255	0.3	0.192 $\pm$ 0.334
	Relaxed	8	0.3	0.162 $\pm$ 0.29	-0.6	-0.444 $\pm$ 0.296

Table 8.1: The table shows the number of participants who recognised the intended affect in each version of the clap and look-around gestures. The cells are coloured cyan for recognition rates between $50\% - 75\%$. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

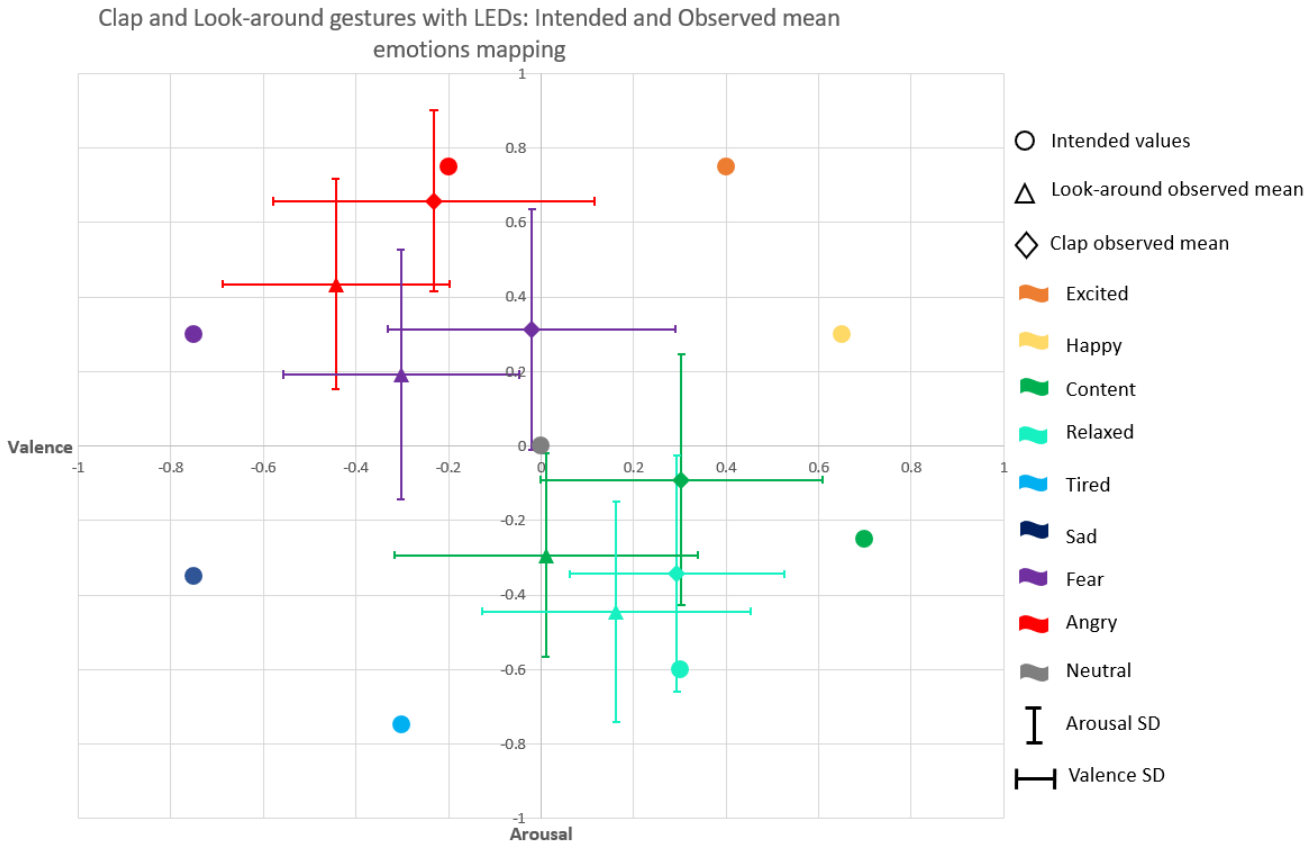


Figure 8.1: Intended values and observed means of affects for the clap and look-around gestures, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

8.1.2. Nod-yes and *These* gestures

Both the gestures have a notable improvement in recognition rate of *anger* as seen from table 8.2. Figure 8.2 shows that mean ratings of *anger, content* and *relaxed* are mapped into the intended quadrants, with *anger* and *relaxed* close to the intended points. However, *fear* does not fall in the correct quadrant in either of the gestures. Though *content* is in the intended quadrant, the mean ratings are still considerably far from the intended valence and arousal values.

Gesture	Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Nod-yes	Anger	10	-0.2	-0.182 $\pm$ 0.354	0.75	0.586 $\pm$ 0.236
	Content	13	0.7	0.283 $\pm$ 0.189	-0.25	-0.162 $\pm$ 0.334
	Fear	4	-0.75	0.081 $\pm$ 0.354	0.3	0.061 $\pm$ 0.386
	Relaxed	8	0.3	0.303 $\pm$ 0.226	-0.6	-0.313 $\pm$ 0.353
These	Anger	17	-0.2	-0.364 $\pm$ 0.268	0.75	0.414 $\pm$ 0.289
	Content	2	0.7	0.05 $\pm$ 0.434	-0.25	-0.081 $\pm$ 0.408
	Fear	3	-0.75	-0.172 $\pm$ 0.426	0.3	-0.374 $\pm$ 0.417
	Relaxed	9	0.3	0.141 $\pm$ 0.312	-0.6	-0.465 $\pm$ 0.333

Table 8.2: The table shows the number of participants who recognised the intended affect in each version of nod-yes and *these* gestures. The cells are coloured cyan for recognition rates between $50\% - 75\%$. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

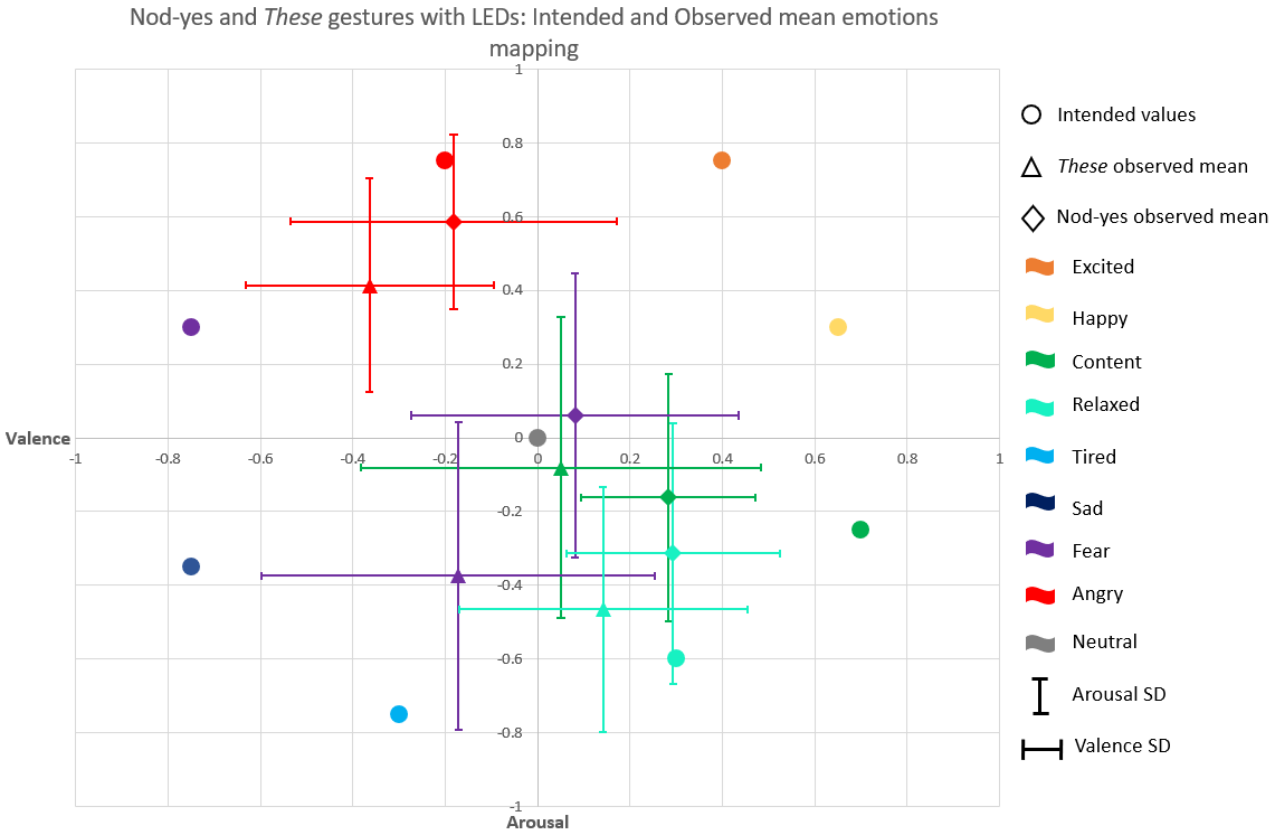


Figure 8.2: Intended values and observed means of affects for the nod-yes and *these* gestures, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

8.1.3. Pointing and Handshake gestures

Table 8.3 shows that both pointing and handshake gestures have significant improvement in recognition rates of *anger* and *fear*. Similar to previous gestures, mean ratings of *anger, content* and *relaxed* fall into the correct quadrants. It can also be seen from figure 8.3 that *anger* and *relaxed* are close to the intended points while *content* is moderately far. *Fear* is mapped to the intended quadrant for pointing gesture but not for handshake gesture.

Gesture	Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
Pointing	Anger	21	-0.2	-0.384 $\pm$ 0.4	0.75	0.556 $\pm$ 0.198
	Content	7	0.7	0.111 $\pm$ 0.35	-0.25	-0.232 $\pm$ 0.395
	Fear	9	-0.75	-0.293 $\pm$ 0.38	0.3	0.202 $\pm$ 0.363
	Relaxed	8	0.3	0.162 $\pm$ 0.334	-0.6	-0.424 $\pm$ 0.375
Handshake	Anger	20	-0.2	-0.353 $\pm$ 0.35	0.75	0.586 $\pm$ 0.264
	Content	5	0.7	0.091 $\pm$ 0.315	-0.25	-0.202 $\pm$ 0.456
	Fear	4	-0.75	-0.051 $\pm$ 0.364	0.3	-0.192 $\pm$ 0.471
	Relaxed	8	0.3	0.222 $\pm$ 0.259	-0.6	-0.354 $\pm$ 0.311

Table 8.3: The table shows the number of participants who recognised the intended affect in each version of the pointing and handshake gestures. The cells are coloured cyan for recognition rates between $50\% - 75\%$. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

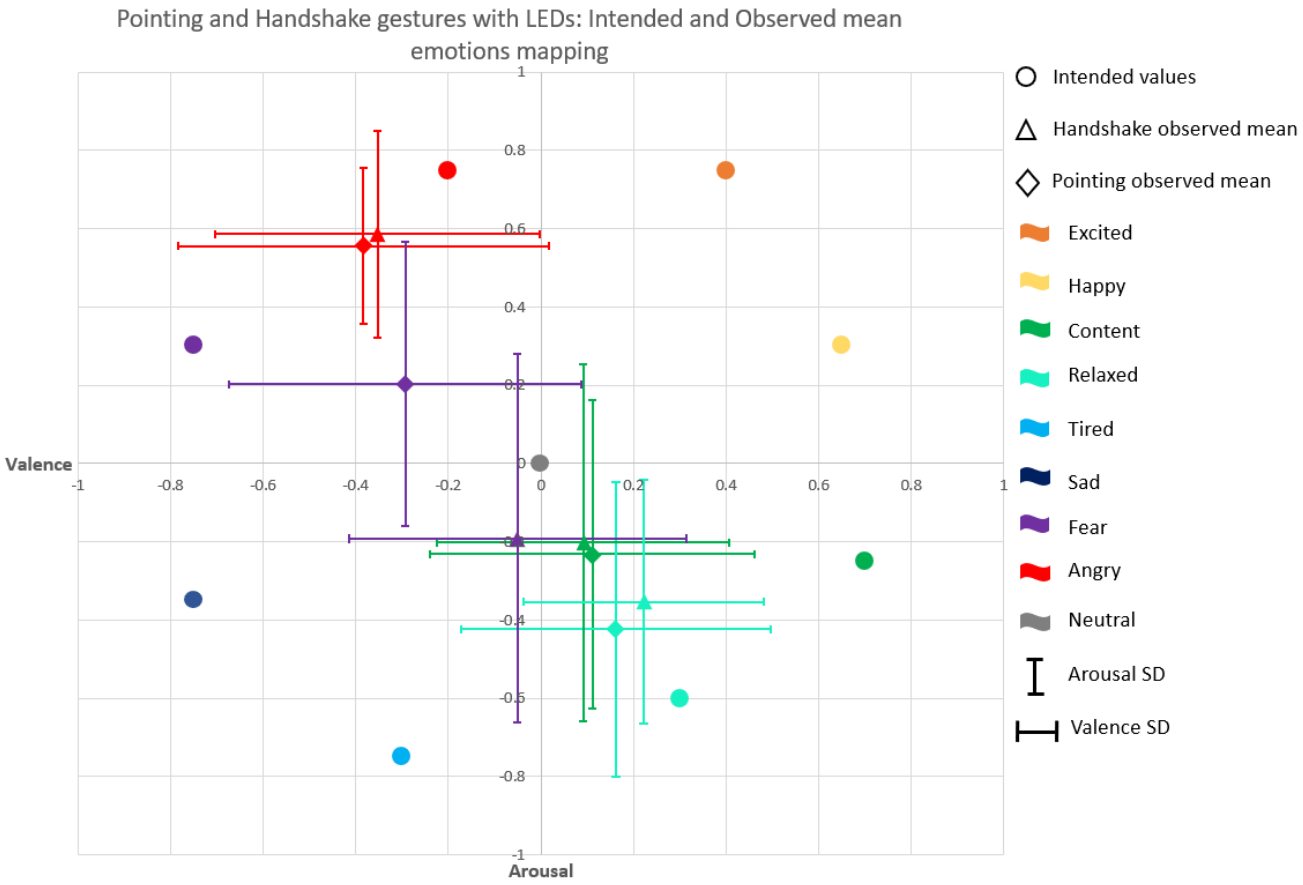


Figure 8.3: Intended values and observed means of affects for the pointing and handshake gestures, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

8.1.4. This-or-that and Wave gestures

The recognition rate of *anger* is moderately high for both gestures as seen in table 8.4. The mean ratings of *anger, fear* and *relaxed* fall into the intended quadrants as seen in figure 8.4. *Anger* and *relaxed* are close to the intended points whereas *content* and *fear* are moderately far.

Gesture	Intended affect	# recognised	Intended valence	Obs. valence $\mu \pm SD$	Intended arousal	Obs. arousal $\mu \pm SD$
This-or-that	Anger	19	-0.2	-0.242 $\pm$ 0.326	0.75	0.586 $\pm$ 0.25
	Content	7	0.7	-0.02 $\pm$ 0.235	-0.25	-0.343 $\pm$ 0.358
	Fear	3	-0.75	-0.071 $\pm$ 0.298	0.3	0.081 $\pm$ 0.354
	Relaxed	8	0.3	0.152 $\pm$ 0.278	-0.6	-0.485 $\pm$ 0.301
Wave	Anger	18	-0.2	-0.192 $\pm$ 0.373	0.75	0.606 $\pm$ 0.227
	Content	8	0.7	0.253 $\pm$ 0.264	-0.25	-0.061 $\pm$ 0.348
	Fear	3	-0.75	-0.02 $\pm$ 0.353	0.3	0.212 $\pm$ 0.321
	Relaxed	8	0.3	0.222 $\pm$ 0.308	-0.6	-0.354 $\pm$ 0.381

Table 8.4: The table shows the number of participants who recognised the intended affect in each version of this-or-that and wave gestures. The cells are coloured cyan for recognition rates between $50\% - 75\%$. It also shows the mean (μ) and standard deviation (SD) of valence and arousal ratings.

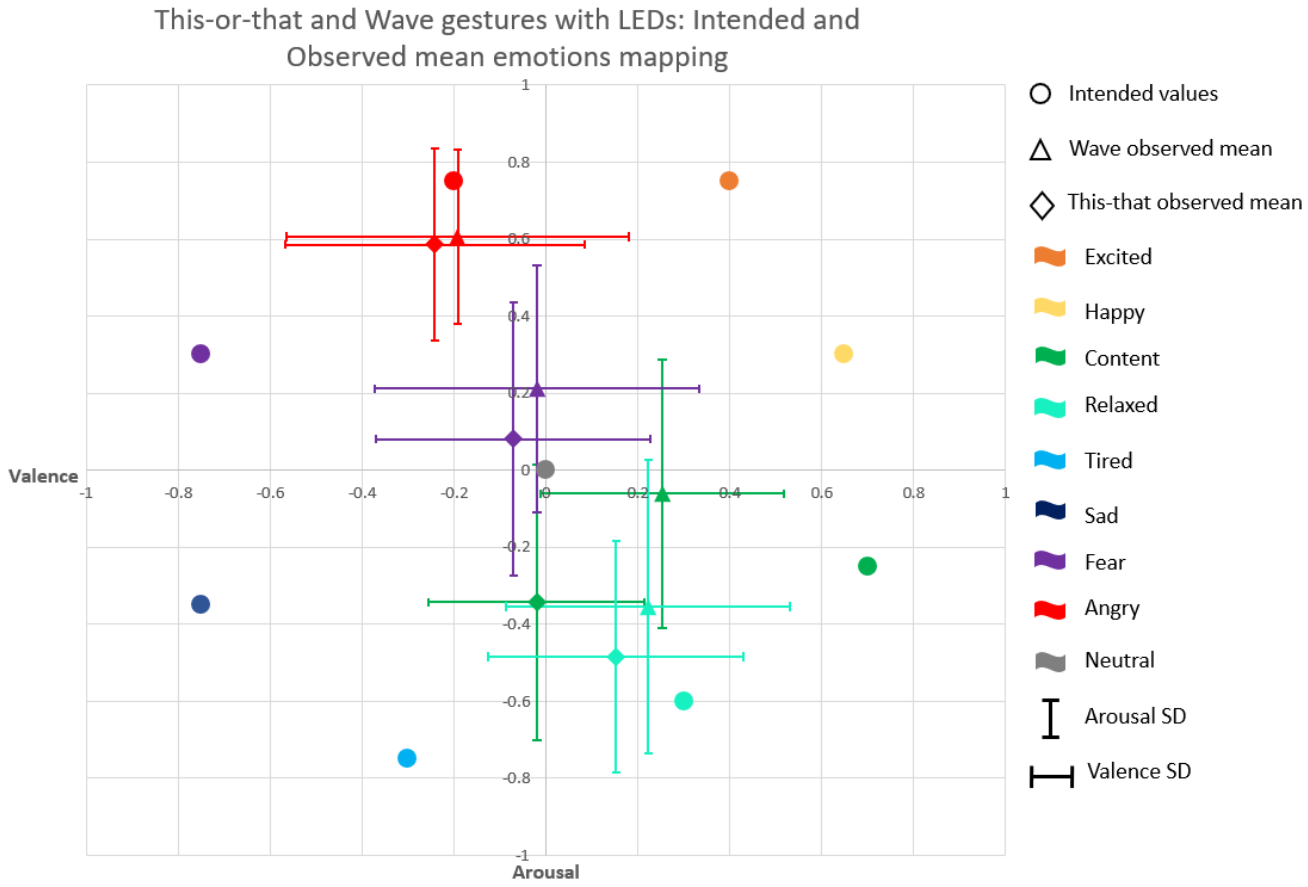


Figure 8.4: Intended values and observed means of affects for the wave and this-or-that gestures, mapped onto the valence-arousal space. The mean affect points also show the standard deviation along valence and arousal.

8.2. Key results

The following points highlight the observed trends while combining LED patterns with motion and body language features.

1. The mean valence and arousal ratings of *anger* and *relaxed* are mapped into the intended quadrants in all gestures. Additionally, these affects are rated close to the intended values.

2. The mean ratings of *fear* fall into the intended quadrant for the majority of gestures (5

out of 8). Though the observed means have moved closer to the intended points, the distance is still considerable.

3. The mean ratings of *content* are in the intended quadrant for 7 gestures. Green LED patterns seems to improve the valence ratings of *content* and *relaxed*. However, this does not catapult the valence ratings to high and hence, *content* is moderately far from the intended values.

8.3. Phase 3

After phase 2, *fear* and *content* still have low recognition rates. The valence-arousal ratings are also considerably far from the intended values. As a last resort, the emotion-specific pose repertoires are considered, and the constraint on task interruption is relaxed. As discussed previously in chapter 6, this phase focused on the ease of recognising specific emotions. Since *content* lacks well-known key poses, only *fear* was tested. A small study involving 10 participants was conducted to evaluate the ease of recognition.

First, the participants were asked to describe the video of the NAO robot enacting the fear pose illustrated in figure 5.4. Most of the participants associated the pose with dodging a hit or shielding its face. Some participants (3 out of 10) used the words 'scared' or 'afraid' in their responses.

Next, the participants had to name an emotion they associate with the pose repertoire. Most participants (9 out of 10) responded with labels such as ' afraid', ' fear' or ' scared'. One participant labelled it as ' anticipation'.

In the last question, the participants had to pick an emotion from the given options. All participants chose ' fear' as the portrayed emotion.

9

Conclusion

Experiments were conducted to test the hypotheses formulated in Chapter 1. The results of these experiments which were conducted in multiple phases are presented in chapters 7 and 8. The key results were also highlighted in the respective chapter. This chapter revisits the hypotheses, discusses the results and other observed trends and examines the consistencies of these results with other related works. It also discusses the contributions, limitations and prospects of the proposed framework.

9.1. Discussions

The phase 1 experiments tested the hypotheses about affect expression capabilities of a model that uses only motion and body language modulations.

1(a). What are the motion and body language features and the associated operators that can be used in a parametric affect expression framework?

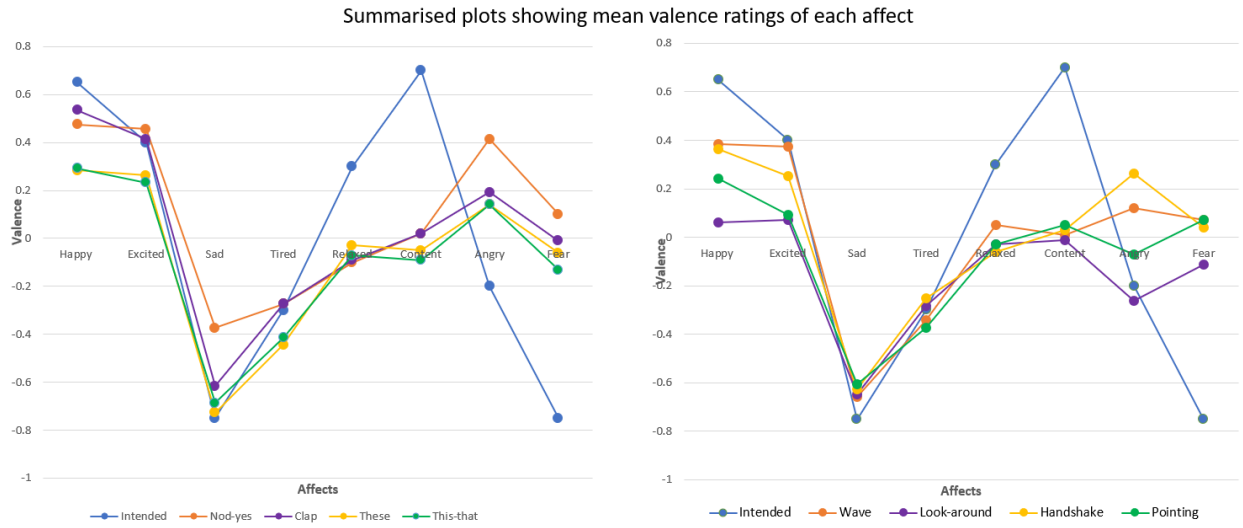


Figure 9.1: Plots summarising the mean valence score (phase 1) of each of the affects. Each line corresponds to a gesture.

After examining various works which use motion and body language features to express affect, 2 valence-oriented features *(amplitude, vertical head pose)* and 3 arousal-oriented features *(speed, repetition, stance)* were chosen. The phase 1 data show that the mean arousal ratings are often close to the intended values for all 8 affects except fear. This indicates

Figure 9.2: Plots summarising the mean arousal score (phase 1) of each of the affects. Each line corresponds to a gesture.

that the chosen features are successful in portraying arousal. The experiments did not examine the impact of individual features on expressing valence or arousal, but many studies like [2, 21, 30, 32, 34, 35] show speed to be a prominent indicator of energy or arousal. The valence ratings of some affects were farther from the intended values than expected. A plausible reason could be the lack of dominant valence-oriented features among the chosen features. While [34, 35] suggests that amplitude is correlated with valence, [5] found correlations between amplitude and arousal. Similarly, [13, 30, 34] suggests vertical head poses for distinguishing affects in the first and third quadrant of the valence-arousal space. However, [1, 13] provides conflicting evidence regarding head poses of anger and fear. Additionally, we did not find any literature which suggests head poses for expressing *content* and *relaxed*. Hence, the head poses were used only for affects in quadrant 1 and 3. The valence ratings of these affects were quite close to the intended values, which demonstrates that the chosen features can express valence to some extent, but are not sufficient for expressing all affects.

Figure 9.1 and 9.2 shows a summarised view of the mean valence and arousal ratings. All affects, except fear, follow the intended arousal line. This demonstrates that motion and body language model successfully portrays the arousal of the affects. In the case of valence, the intended line is followed only by half of the affects, i.e. for happy, excited, sad, and tired. Other affects show significant deviations from the intended line.

As seen in many of the related works, some affects are easier to express than others. Additionally, gestures like nodding involves motion along a single joint (head pitch) which is also a body language feature. This may have hampered the expressive capability of the model in the specific case of nodding. Hence it is interesting to evaluate which affects are expressed sufficiently well by the model.

1(b). Which affects expressed by the motion and body language model are recognisable and distinguishable ?

Excited, happy and *sad* were hypothesised to have a high recognition rate and be distinguishable in their valence-arousal ratings. Sad and excited had high recognition rates in the majority of gestures and were rated close to the intended values. Happy had a moderate recognition rate and was slightly farther from the intended valence-arousal values. However, this was still in the acceptable range and consistent with the values of happy in [27, 33]. As expected, *tired* had a moderate recognition rate. Interestingly, the valence-arousal ratings of tired were quite close to the intended values. The mean ratings of anger, relaxed, content and fear were mostly not in the expected quadrant and had low recognition rates.

This is consistent with [30] which also had difficulty in expressing affects in the second and fourth quadrant. Anger was often confused as excited or happy, which was also reflected in the valence-arousal ratings. This is consistent with studies like [5, 9, 35] which also had difficulty in expressing anger due to such confusions. A statistical test was run on the valence and arousal ratings provided by the participants, to determine whether the affects were distinguishable. Sad and tired were often distinguishable along both valence and arousal, whereas happy and excited were mostly distinguishable only by arousal. Hence, the proposed motion and body language model can express 4 affects (excited, happy, sad, tired) that are distinguished by the users.

A proposed enhancement was employing the LED channel to improve the perception of various affects. In phase 2 of the experiments, the hypotheses regarding the addition of LED patterns were tested.

2(a). What are the colours and patterns which can be used for expressing various affects?

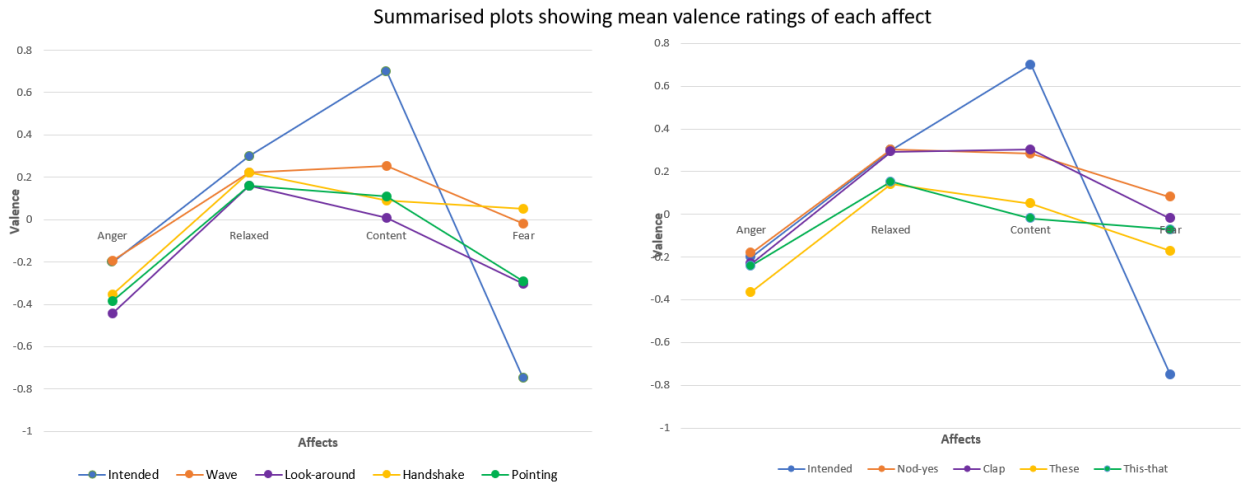


Figure 9.3: Plots summarising the mean valence score (phase 2) of each of the affects. Each line corresponds to a gesture.

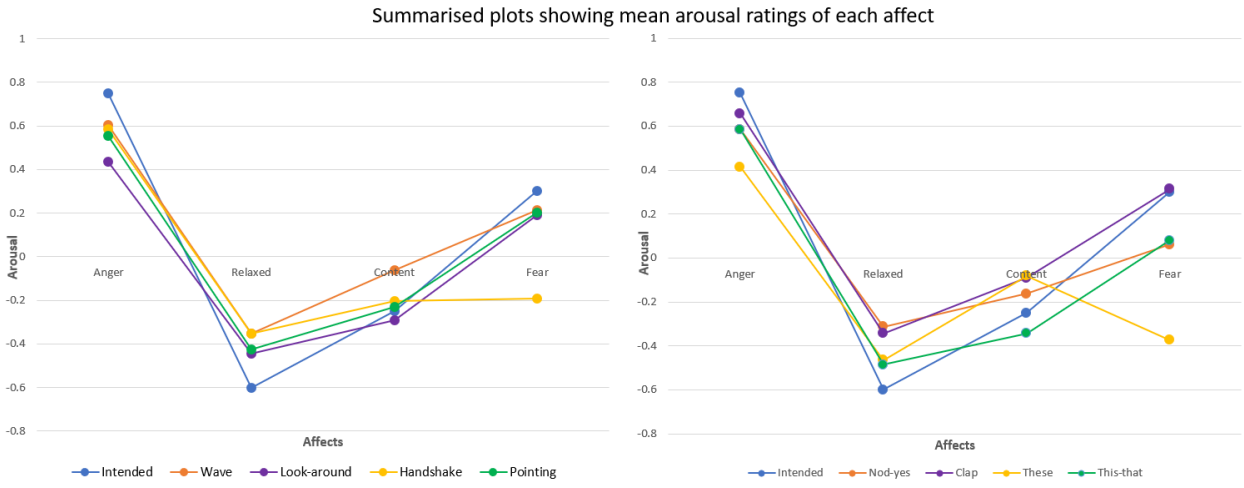


Figure 9.4: Plots summarising the mean arousal score (phase 2) of each of the affects. Each line corresponds to a gesture.

Hues of red were associated with high arousal and hues of blue were associated with low arousal, which is on par with the expectations. It was also observed that hues of green tend

to express positive affect. As observed in [14, 28], the blinking pattern (rise and fall) and frequency portrayed arousal. In the majority of gestures, the mean valence-arousal ratings of anger, fear, relaxed and content were mapped onto the intended quadrant. Thus, the lack of a strong valence indicator is masked by adding LED patterns.

Figure 9.3 and 9.4 shows a summarised view of the mean valence and arousal ratings of phase 2. All affects (including fear) follow the intended arousal line. This demonstrates that adding LED patterns improved the portrayal of arousal. In the case of valence, the intended line is followed only by anger and relaxed. Fear and content show significant deviations from the intended line.

Adding LED patterns would increase the recognition of the expressed affects. Similar to phase 1, this hypothesis needs to be evaluated.

2(b). What additional affects are perceived by incorporating LED patterns?

Adding LED patterns improved the recognition rates of anger and fear. For *anger*, the rates improved to a moderate level, though a higher rate was expected. Unlike the observation in [28], fear was not recognised well. The mean valence-arousal ratings of *anger* and *relaxed* were quite close to the intended values. In many cases, arousal ratings of fear and content were close to the intended values. Hues of green seem to indicate a positive valence, but the mean valence scores of content and relaxed were similar. This contrasts [10], which suggests that different hues of green represent different levels of relaxation. Unlike our experiments, [10] used mobile phones for their study. The difference in the platform could be the reason for different observations. The results suggest that adding LED patterns allow users to recognise two more affects (anger, relaxed), increasing the total number of perceived affects to six.

As discussed in chapter 1, emotion-specific poses are used to express affects which were not perceived using the previous models.

3. What are the emotion specific pose repertoires that can be added to increase the perceivable affects?

Fear and *content* were not recognised in the previous phases. Studies like [3, 7] designed key poses for fear involving averted gaze and hands covering the face. Since *content* does not have a well-known key pose, this phase tested only fear. A small study revealed that 90% of the participants associated fear with the presented key pose. All the participants picked fear as the portrayed emotion when they had to choose from the given list. Hence, fear is easily recognised through a pose repertoire.

9.2. Contributions

This thesis focused on three familiar techniques for expressing affect and set out to identify the range of affects that can be expressed using these techniques. The contributions of this thesis are as follows:

1. Mathematical models: This thesis investigated the three affect expression techniques and proposed a few features and operators for implementing them. It formulated mathematical models for these features and also defined operators to modulate them. The parameters of these models define the transformations done to the input gesture. Different frameworks involving different transformations can be developed by changing the parameters of the models.

2. Generic affect expression framework: The valence-arousal affect space is continuous. Hence instead of an emotion-specific framework which renders fixed discrete affects, the aim was to create a framework which can render any affect specified as a point in

the affect space. Since the foundation of the framework was the mathematical models which can be controlled by parameters, these parameters were calculated solely based on valence-arousal values of the input affect. Additionally, the framework is not robot specific and hence the affective gesture output can be used in any humanoid robot.

3. Complexity of expression: The three techniques can be viewed as layers of complexity in the framework. Additional techniques were used only when the affect was not expressible by the previous model. For example, affects like sad, excited etc. can be seen as less complex affects to express since they were perceived as intended by using just the motion and body language model. On the other hand, expressing fear requires additional emotion-specific pose repertoire, which makes it a very complex affect to express.

4. Decision on when to use a model: This thesis also demonstrates a systematic method to determine when to use a model for affect expression. For example, anger has been shown by many studies to be a difficult affect to express. We systematically determined through experiments the best expression model for each of the 8 affects. In some case like tired, even if the recognition rate is moderate, the perceived valence and arousal are as intended.

5. Extensive experiments: The experiments included 8 gestures and up to 8 affects in each phase. The gesture list covered movements across all joints in the head and arms. The affects were chosen from all quadrants of valence-arousal space and represented 5 levels of valence and arousal. Phase 1 had 72 test videos and phase 2 had 32 videos. In addition to the affect label data, we also collected and analysed valence and arousal ratings. Due to the large volume of data involved, the experiments were conducted online.

9.3. Limitations

Though the framework yielded good results and provided evidence supporting the hypotheses, there is still scope for improvement. A few aspects can be investigated further to improve the framework.

First, as seen in the phase 1 results, the current framework lacks a strong indicator of valence. Though this was ameliorated by adding LED patterns, the valence ratings of a few affects like fear and content are still farther than intended. Hence, more valence-oriented features need to be explored.

Second, the proposed models use simple modulations. All features are classified as valence-oriented or arousal-oriented. Depending on this classification, the underlying modulations rely solely on either valence or arousal. But some studies [5, 21, 34] have shown that features like amplitude, repetition, stance etc. influence both perceived valence and arousal. Hence, these features should be further investigated to develop complex modulations which depend on both valence and arousal.

Third, some of the parameter values used to instantiate the models were determined empirically from illustrations in literature. The data presented in chapters 7 and 8 can be examined to determine which of these values need further calibration. An extensive study involving human participants could determine more realistic values for these parameters. This could also help in fine-tuning of the parameters for improving the results.

Fourth, the experiments were conducted online using test videos. Using live setting instead of recorded videos could alter the perceptions slightly. Some features might receive more attention than others in a live setting. The videos could capture the colours completely. There is a difference in the LED captured by a normal digital camera and viewed directly by our eyes. Though this is a limitation of the camera, it could have some influence on phase 2 results. The phase 2 experiments may yield better results in a live setting.

Fifth, the impact of adding LEDs to affects like sad need to be investigated. These affects were already perceived using the motion and body language model. Though no degradation in the results is expected, it still needs to be verified.

Lastly, we did not find an effective method for expressing the emotion *content*. There aren't many studies which focus on expressing content. Solving the lack of strong valence indicators may resolve this issue as well. The improvements observed by adding a green LED pattern is promising, but more features are required to increase the recognition rate of content.

9.4. Future works

While we already discussed some areas of improvement in the previous section, this section suggests prospects for enhancing and extending the framework.

First, it is crucial to study the effect of gender, age or cultural diversity on the results. The framework aims to facilitate and improve human-computer interactions, which makes it necessary to investigate and explain any difference in perception or responses among diverse groups. If the framework is deployed in an application targeting a particular demographic group, it is recommended to tailor the framework and its parameters through focused experiments.

Second, experiments can be conducted to evaluate the recognition of affects in a given context. All the experiments conducted as part of this thesis were context-free, i.e. the participants viewed the performance of the robot without any context. It would be interesting to study the effectiveness of the framework in narration or story-telling, where each gesture has an associated context. While evaluating the gesture videos, the participants knew what their task was and focused on figuring out the expressed affect. Thus, some of the affects may have been recognised because of high attentiveness, and may not be noticed in a natural setting with moderate to low cognitive load. A narration or story-telling based experiment would provide such an environment, where the participants may divide their attention among other aspects like the current context, the plot of the story etc.

Third, many studies use a 3-D affect space involving valence, arousal and dominance axes rather than a 2-D affect space. Features like gaze, approach/avoidance, etc. could portray dominance. Adding features which are strong indicators of dominance would improve the framework and may increase the range of perceivable affects.

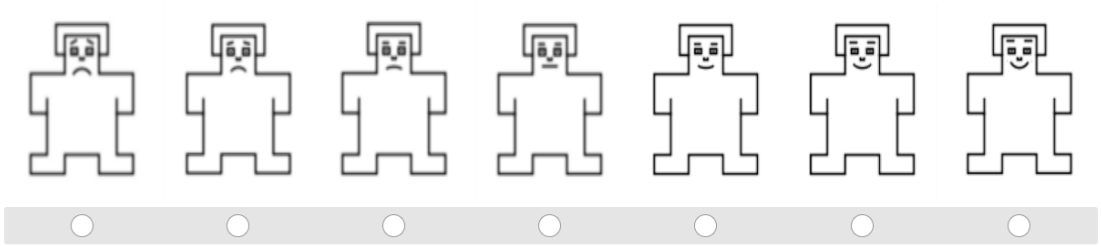
Lastly, the framework could be improved by employing additional channels for expressing affect. This thesis already demonstrated that using an additional channel (LEDs) along with motion and body language features improves the framework. Similarly, speech or voice channel can be explored to improve the expressive capabilities of the framework without disrupting the task being performed.

A

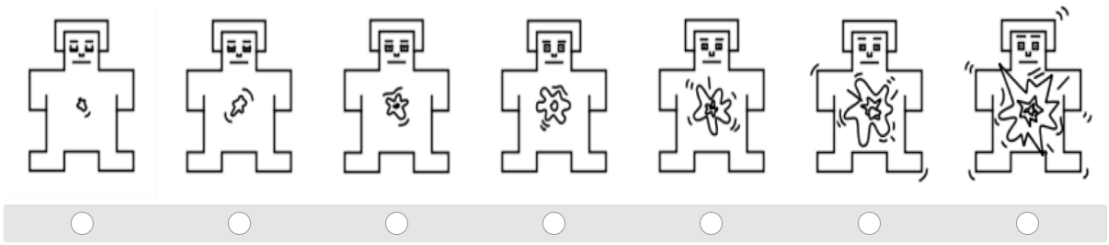
SAM questionnaire

SAM questionnaire was used to obtain valence and arousal ratings from the participants. We used the 7-point version of the scale used in [4]. SAM consists of rating scales for valence, arousal and dominance but for the experiments only valence and arousal scales were used. The SAM questionnaire was presented as:

What was the **valence** (**negative** vs **positive** emotion) of the robot?



What was the **arousal** (**activation** or **energy**) of the robot?



The participants were given a small explanation and sufficient examples to understand the SAM scales. The following information were given to the participants in the beginning of the experiment.

Valence (pleasure level) shows positive versus negative emotional state.
Low valence e.g.: sad, fear
Moderate valence e.g.: angry, neutral, relaxed
High valence e.g.: satisfied, happy

Arousal (energy) shows level of mental alertness or physical activity.
Low arousal e.g.: tired, relaxed
Moderate arousal e.g.: sad, neutral, happy
High arousal e.g.: excited, angry

B

Demography

The following tables list the demographic details of participants for each gesture.

Gesture	Gender	Age	Country
Wave	M - 16, F - 15, O - 1, N - 1	21 - 60 yrs (μ = 35.8 yrs)	U.S.A - 29, U.K - 1, Canada - 3
Look-around	M - 20, F - 13	19 - 55 yrs (μ = 32.3 yrs)	U.S.A - 32, U.K - 1
Handshake	M - 11, F - 22	20 - 59 yrs (μ = 36.8 yrs)	U.S.A - 30, Canada - 3
Nod-yes	M - 16, F - 17	22 - 53 yrs (μ = 35.7 yrs)	U.S.A - 30, U.K - 2, Canada - 1
Clap	M - 9, F - 24	21 - 64 yrs (μ = 43.7 yrs)	U.S.A - 29, U.K - 2, Canada - 2
Pointing	M - 12, F - 21	19 - 58 yrs (μ = 35.9 yrs)	U.S.A - 33
These	M - 13, F - 20	18 - 52 yrs (μ = 36.6 yrs)	U.S.A - 31, U.K - 1, Canada - 1
This-that	M - 14, F - 19	22 - 61 yrs (μ = 35.0 yrs)	U.S.A - 31, Canada - 2

Table B.1: Demographic details of participants - phase 1. M denotes male, F denotes female, O denotes other and N denotes prefer not to say.

Gesture	Gender	Age	Country
Clap & Look-around	M - 23, F - 10	22 - 55 yrs (μ = 33.4 yrs)	U.S.A - 33
These & Nod-yes	M - 17, F - 16	20 - 55 yrs (μ = 33.6 yrs)	U.S.A - 30, Canada - 3
Handshake & Pointing	M - 19, F - 14	19 - 68 yrs (μ = 36.8 yrs)	U.S.A - 33
Wave & This-that	M - 17, F - 15, Other - 1	22 - 67 yrs (μ = 35.7 yrs)	U.S.A - 29, U.K - 1, Canada - 3

Table B.2: Demographic details of participants - phase 2. M denotes male and F denotes female.

C

Emotion Recognition data - Phase 1

In the phase 1 experiment, the participants were asked to choose the emotion they thought the robot expressed. The participants had 9 options to choose from: excited, happy, content, relaxed, tired, sad, fear, angry and neutral. It is natural that the emotion perceived by the participants is not always the intended emotion. The following tables show the emotion label responses each gesture video received.

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	25	1	0	0	0	0	0	0	7
Happy	10	17	0	0	1	0	0	2	3
Excited	0	4	29	0	0	0	0	0	0
Anger	0	8	15	3	3	0	0	0	4
Fear	13	3	0	0	0	3	2	7	5
Sad	0	0	0	0	0	28	5	0	0
Tired	0	0	0	0	0	11	19	3	0
Relaxed	4	0	0	0	0	2	14	9	4
Content	17	0	0	1	0	2	2	5	6

Table C.1: The perceived emotions and the number of participants who chose the emotions for the various versions of wave gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	23	0	1	0	4	1	0	1	3
Happy	4	9	7	2	6	1	0	1	3
Excited	0	3	18	3	9	0	0	0	0
Anger	4	0	9	12	8	0	0	0	0
Fear	7	2	6	2	5	3	2	2	4
Sad	0	0	0	0	1	27	2	3	0
Tired	3	0	0	0	2	12	13	3	0
Relaxed	4	1	0	0	2	6	9	7	4
Content	16	1	0	0	2	3	3	4	4

Table C.2: The perceived emotions and the number of participants who chose the emotions for the various versions of look-around gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	26	1	0	0	0	0	0	6	0
Happy	3	11	7	1	1	0	0	2	8
Excited	1	5	24	3	0	0	0	0	0
Anger	3	7	16	2	0	0	0	0	5
Fear	18	1	0	0	0	2	4	4	4
Sad	2	0	0	0	5	26	0	0	0
Tired	1	0	0	0	0	9	17	4	2
Relaxed	6	1	0	0	5	2	9	4	6
Content	15	2	2	0	1	1	4	4	4

Table C.3: The perceived emotions and the number of participants who chose the emotions for the various versions of handshake gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	15	1	2	0	0	1	0	3	11
Happy	0	15	13	1	1	0	1	0	2
Excited	0	4	28	1	0	0	0	0	0
Anger	1	8	22	0	0	1	0	0	1
Fear	9	4	2	0	0	1	2	3	12
Sad	3	0	1	0	3	26	0	0	0
Tired	1	0	0	0	2	12	14	1	3
Relaxed	3	2	2	1	3	9	5	7	1
Content	5	2	1	0	0	3	7	9	6

Table C.4: The perceived emotions and the number of participants who chose the emotions for the various versions of nod-yes gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	18	9	3	0	0	0	0	0	3
Happy	2	22	3	0	0	0	0	3	3
Excited	0	2	25	6	0	0	0	0	0
Anger	0	11	15	5	0	0	0	0	2
Fear	6	6	1	0	2	2	2	6	8
Sad	0	0	0	0	0	25	8	0	0
Tired	1	0	0	0	1	17	9	0	5
Relaxed	1	0	0	0	0	6	14	10	2
Content	5	2	0	0	0	2	5	10	9

Table C.5: The perceived emotions and the number of participants who chose the emotions for the various versions of clap gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	22	3	0	2	1	0	1	1	3
Happy	3	9	5	7	0	0	0	2	7
Excited	0	4	18	9	1	0	0	0	1
Anger	1	3	15	11	2	0	0	0	1
Fear	7	4	6	1	2	1	3	6	3
Sad	0	0	0	0	4	22	6	0	1
Tired	2	0	0	2	3	11	14	0	1
Relaxed	5	1	1	0	3	5	12	3	3
Content	11	1	0	1	0	3	4	10	3

Table C.6: The perceived emotions and the number of participants who chose the emotions for the various versions of pointing gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	23	0	1	0	0	4	1	0	4
Happy	3	10	3	3	1	0	1	4	8
Excited	0	3	26	2	0	0	0	1	1
Anger	6	6	12	8	0	0	0	0	1
Fear	11	3	4	0	0	3	6	4	2
Sad	0	0	0	0	0	30	3	0	0
Tired	1	0	0	2	2	17	10	0	1
Relaxed	2	0	0	0	4	7	7	8	5
Content	8	2	2	2	3	0	7	3	6

Table C.7: The perceived emotions and the number of participants who chose the emotions for the various versions of *these* gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Neutral	22	0	0	0	1	3	0	6	1
Happy	3	17	4	3	1	0	0	0	5
Excited	0	4	25	3	1	0	0	0	0
Anger	2	9	12	5	0	0	0	0	5
Fear	9	2	0	0	1	7	2	6	6
Sad	0	0	0	0	1	29	2	0	1
Tired	1	0	0	0	1	14	13	4	0
Relaxed	12	2	0	0	0	6	8	2	3
Content	15	0	0	0	0	4	5	5	4

Table C.8: The perceived emotions and the number of participants who chose the emotions for the various versions of this-or-that gesture. The rows are coloured green for identification rates ≥ 75% and cyan for rates between 50% – 75%

D

Emotion Recognition data - Phase 2

Phase 2 followed the same questionnaire format as phase 1. The following tables show the emotion label responses each gesture video received. Note that only 4 affects were tested in this phase.

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	1	3	11	18	0	0	0	0	0
Fear	10	8	3	1	3	3	0	0	5
Relaxed	7	9	0	0	1	3	2	8	3
Content	7	10	4	1	1	1	1	0	8

Table D.1: The perceived emotions and the number of participants who chose the emotions for the various versions of wave gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	2	1	1	19	9	1	0	0	0
Fear	7	2	1	4	8	8	1	2	0
Relaxed	7	2	1	0	3	4	5	8	3
Content	12	2	1	1	2	5	5	1	4

Table D.2: The perceived emotions and the number of participants who chose the emotions for the various versions of look-around gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	0	1	7	20	3	0	0	1	1
Fear	7	2	3	1	4	2	2	6	6
Relaxed	9	5	2	0	0	0	1	8	8
Content	7	6	5	2	1	1	1	5	5

Table D.3: The perceived emotions and the number of participants who chose the emotions for the various versions of handshake gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	3	3	10	10	6	0	0	1	0
Fear	5	8	2	2	4	0	0	5	7
Relaxed	3	6	1	0	0	1	2	8	12
Content	3	14	2	0	1	0	0	0	13

Table D.4: The perceived emotions and the number of participants who chose the emotions for the various versions of nod-yes gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	0	4	10	18	1	0	0	0	0
Fear	6	6	10	2	7	0	1	0	1
Relaxed	2	8	6	0	1	0	4	5	7
Content	3	9	9	0	1	0	2	1	8

Table D.5: The perceived emotions and the number of participants who chose the emotions for the various versions of clap gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	0	1	5	21	5	1	0	0	0
Fear	3	0	2	10	9	1	2	1	5
Relaxed	5	3	2	0	4	2	5	8	4
Content	7	5	1	1	4	3	2	3	7

Table D.6: The perceived emotions and the number of participants who chose the emotions for the various versions of pointing gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	5	0	3	17	4	4	0	0	0
Fear	7	0	2	1	3	8	9	0	3
Relaxed	4	2	1	0	1	4	8	9	4
Content	5	5	3	2	2	3	6	5	2

Table D.7: The perceived emotions and the number of participants who chose the emotions for the various versions of *these* gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

	Perceived affect								
Intended affect	Neutral	Happy	Excited	Anger	Fear	Sad	Tired	Relaxed	Content
Anger	3	0	7	19	2	0	1	0	1
Fear	10	0	4	1	3	4	3	6	2
Relaxed	11	1	0	0	1	4	5	8	3
Content	11	0	0	1	0	4	5	5	7

Table D.8: The perceived emotions and the number of participants who chose the emotions for the various versions of this-or-that gesture. The rows are coloured cyan for identification rates between $50\% - 75\%$

Bibliography

[1] Reginald B Adams Jr and Robert E Kleck. Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1):3, 2005.

[2] Kenji Amaya, Armin Bruderlin, and Tom Calvert. Emotion from motion. In *Graphics interface*, volume 96, pages 222–229. Toronto, Canada, 1996.

[3] Aryel Beck, Lola Cañamero, and Kim A Bard. Towards an affect space for robots to display emotional body language. In *19th International symposium in robot and human interactive communication*, pages 464–469. IEEE, 2010.

[4] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[5] Ginevra Castellano, Santiago D Villalba, and Antonio Camurri. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 71–82. Springer, 2007.

[6] Gabriel Castillo and Michael Neff. What do we express without knowing?: Emotion in gesture. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pages 702–710, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-1-4503-6309-9. URL `http://dl.acm.org/citation.cfm?id=3306127.3331759`.

[7] Iris Cohen, Rosemarijn Looije, and Mark A Neerincx. Child's recognition of emotions in robot's face and body. In *Proceedings of the 6th international conference on Human-robot interaction*, pages 123–124. ACM, 2011.

[8] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[9] Elizabeth A Crane and M Melissa Gross. Effort-shape characteristics of emotion-related body movement. *Journal of Nonverbal Behavior*, 37(2):91–105, 2013.

[10] Petra Fagerberg, Anna Ståhl, and Kristina Höök. emoto: emotionally engaging interaction. *Personal and Ubiquitous Computing*, 8(5):377–381, 2004.

[11] Alan Hanjalic. Extracting moods from pictures and sounds: Towards truly personalized tv. *IEEE Signal Processing Magazine*, 23(2):90–100, 2006.

[12] Kazuko Itoh, Hiroyasu Miwa, Yuko Nukariya, Massimiliano Zecca, Hideaki Takanobu, Stefano Roccella, Maria Chiara Carrozza, Paolo Dario, and Atsuo Takanishi. Mechanisms and functions for a humanoid robot to express human-like emotions. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 4390–4392, 2006.

[13] David O Johnson and Raymond H Cuijpers. Investigating the effect of a humanoid robot's head position on imitating human emotions. *International Journal of Social Robotics*, pages 1–10, 2019.

[14] David O Johnson, Raymond H Cuijpers, and David van der Pol. Imitating human emotions with artificial facial expressions. *International Journal of Social Robotics*, 5(4):503–513, 2013.

[15] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[16] Heather Knight and Reid Simmons. Laban head-motions convey robot state: A call for robot body language. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 2881–2888. IEEE, 2016.

[17] Hun-ok Lim, Akinori Ishii, and Atsuo Takanishi. Emotion-based biped walking. *Robotica*, 22(5):577–586, 2004.

[18] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[19] Kaya Naz and Helena Epps. Relationship between color and emotion: A study of college students. *College Student J*, 38(3):396, 2004.

[20] Matthias Nieuwenhuisen and Sven Behnke. Human-like interaction skills for the mobile communication robot robotinho. *International Journal of Social Robotics*, 5(4):549–561, 2013.

[21] Jekaterina Novikova and Leon Watts. A design model of emotional body expressions in non-humanoid robots. In *Proceedings of the Second International Conference on Human-agent Interaction*, HAI '14, pages 353–360, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3035-0. doi: 10.1145/2658861.2658892. URL http://doi.acm.org/10.1145/2658861.2658892.

[22] Rifca Peters, Joost Broekens, Kangqi Li, and Mark A Neerincx. Robot dominance expression through parameter-based behaviour modulation. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 224–226. ACM, 2019.

[23] Robert Plutchik. Emotions and psychotherapy: A psychoevolutionary perspective. In *Emotion, psychopathology, and psychotherapy*, pages 3–41. Elsevier, 1990.

[24] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[25] Ali-Akbar Samadani, Sarahjane Burton, Rob Gorbet, and Dana Kulic. Laban effort and shape analysis of affective hand and arm movements. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 343–348. IEEE, 2013.

[26] Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81, 1954.

[27] Kai Sun, Junqing Yu, Yue Huang, and Xiaoqiang Hu. An improved valence-arousal emotion space for video affective content representation and recognition. *2009 IEEE International Conference on Multimedia and Expo*, pages 566–569, 2009.

[28] Kazunori Terada, Atsushi Yamauchi, and Akira Ito. Artificial emotion expression for a robot by dynamic color change. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 314–321. IEEE, 2012.

[29] Myrthe Tielman, Mark Neerincx, John-Jules Meyer, and Rosemarijn Looije. Adaptive emotional expression in robot-child interaction. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 407–414. ACM, 2014.

[30] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx. *Mood expression through parameterized functional behavior of robots*. IEEE, 2013.

[31] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx. The relative importance and interrelations between behavior parameters for robots' mood expression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 558–563. IEEE, 2013.

[32] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx. Bodily mood expression: Recognize moods from functional behaviors of humanoid robots. In *International conference on social robotics*, pages 511–520. Springer, 2013.

[33] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx. Effects of bodily mood expression of a robotic teacher on students. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 2614–2620. IEEE, 2014.

[34] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx. Mood contagion of robot body language in human robot interaction. *Autonomous Agents and Multi-Agent Systems*, 29(6):1216–1248, 2015.

[35] Atsushi Yamaguchi, Yoshikazu Yano, Shinji Doki, and Shigeru Okuma. A study of emotional motion description by motion modification rules using adjectival expressions. In *2006 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 2837–2842. IEEE, 2006.

[36] Zhihong Zeng, Jilin Tu, Brian M Pianfetti, and Thomas S Huang. Audio–visual affective expression recognition through multistream fused hmm. *IEEE Transactions on Multimedia*, 10(4):570–577, 2008.