1-1-2018

Developing an Affect-Aware Rear-Projected Robotic Agent

Ali Mollahosseini
University of Denver

Developing an Affect-Aware Rear-Projected Robotic Agent

Abstract

Social (or Sociable) robots are designed to interact with people in a natural and interpersonal manner. They are becoming an integrated part of our daily lives and have achieved positive outcomes in several applications such as education, health care, quality of life, entertainment, etc. Despite significant progress towards the development of realistic social robotic agents, a number of problems remain to be solved. First, current social robots either lack enough ability to have deep social interaction with human, or they are very expensive to build and maintain. Second, current social robots have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. To address these problems, this dissertation presents the development of a low-cost, flexible, affect-aware rear-projected robotic agent (called *ExpressionBot*), that is designed to support verbal and non-verbal communication between the robot and humans, with the goal of closely modeling the dynamics of natural face-to-face communication.

The developed robotic platform uses state-of-the-art character animation technologies to create an animated human face (aka avatar) that is capable of showing facial expressions, realistic eye movement, and accurate visual speech, and then project this avatar onto a face-shaped translucent mask. The mask and the projector are then rigged onto a neck mechanism that can move like a human head. Since an animation is projected onto a mask, the robotic face is highly flexible research tool, mechanically simple, and low-cost to design, build and maintain compared with mechatronic and android faces. The results of our comprehensive Human-Robot Interaction (HRI) studies illustrate the benefits and values of the proposed rear-projected robotic platform over a virtual-agent with the same animation displayed on a 2D computer screen. The results indicate that ExpressionBot is well accepted by users, with some advantages in expressing facial expressions more accurately and perceiving mutual eye gaze contact.

To improve social capabilities of the robot and create an expressive and empathic social agent (affect-aware) which is capable of interpreting users' emotional facial expressions, we developed a new Deep Neural Networks (DNN) architecture for Facial Expression Recognition (FER). The proposed DNN was initially trained on seven well-known publicly available databases, and obtained significantly better than, or comparable to, traditional convolutional neural networks or other state-of-the-art methods in both accuracy and learning time. Since the performance of the automated FER system highly depends on its training data, and the eventual goal of the proposed robotic platform is to interact with users in an uncontrolled environment, a database of facial expressions in the wild (called *AffectNet*) was created by querying emotion-related keywords from different search engines. AffectNet contains more than 1M images with faces and 440,000 manually annotated images with facial expressions, valence, and arousal. Two DNNs were trained on AffectNet to classify the facial expression images and predict the value of valence and arousal. Various evaluation metrics show that our deep neural network approaches trained on AffectNet can perform better than conventional machine learning methods and available off-the-shelf FER systems.

We then integrated this automated FER system into spoken dialog of our robotic platform to extend and enrich the capabilities of ExpressionBot beyond spoken dialog and create an affect-aware robotic agent that can measure and infer users' affect and cognition. Three social/interaction aspects (task engagement, being empathic, and likability of the robot) are measured in an experiment with the affect-aware robotic agent. The results indicate that users rated our affect-aware agent as empathic and likable as a robot in which user's affect is recognized by a human (WoZ).

In summary, this dissertation presents the development and HRI studies of a perceptive, and expressive, conversational, rear-projected, life-like robotic agent (aka ExpressionBot or Ryan) that models natural face-to-face communication between human and emapthic agent. The results of our in-depth human-robot-interaction studies show that this robotic agent can serve as a model for creating the next generation of empathic social robots.

Document Type
Dissertation

Degree Name
Ph.D.

Department
Computer Science and Engineering

First Advisor
Mohammad H. Mahoor, Ph.D.

Second Advisor
Kateri McRae

Third Advisor
Kimon Valavanis

Keywords
Affect perception, Empathic robot, Facial expressions, Human robot interaction, Rear projected robot, Social robots

Subject Categories
Artificial Intelligence and Robotics | Computer Engineering | Computer Sciences | Robotics

DEVELOPING AN AFFECT-AWARE REAR-PROJECTED ROBOTIC AGENT

A Thesis

Presented to

the Faculty of the Daniel Felix Ritchie School of Engineering and Computer Science

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by

Ali Mollahosseini

March 2018

Advisor: Mohammad H. Mahoor, Ph.D.

Author: Ali Mollahosseini
Title: DEVELOPING AN AFFECT-AWARE REAR-PROJECTED ROBOTIC AGENT
Advisor: Mohammad H. Mahoor, Ph.D
Degree Date: March 2018

Abstract

Social (or Sociable) robots are designed to interact with people in a natural and interpersonal manner. They are becoming an integrated part of our daily lives and have achieved positive outcomes in several applications such as education, health care, quality of life, entertainment, etc. Despite significant progress towards the development of realistic social robotic agents, a number of problems remain to be solved. First, current social robots either lack enough ability to have deep social interaction with human, or they are very expensive to build and maintain. Second, current social robots have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. To address these problems, this dissertation presents the development of a low-cost, flexible, affect-aware rear-projected robotic agent (called *ExpressionBot*), that is designed to support verbal and non-verbal communication between the robot and humans, with the goal of closely modeling the dynamics of natural face-to-face communication.

The developed robotic platform uses state-of-the-art character animation technologies to create an animated human face (aka avatar) that is capable of showing facial expressions, realistic eye movement, and accurate visual speech, and then project this avatar onto a face-shaped translucent mask. The mask and the projector are then rigged onto a neck mechanism that can move like a human head. Since an animation is projected onto a mask, the robotic face is highly flexible research tool, mechanically simple, and low-cost to design, build and maintain compared with mechatronic and android faces. The results of our comprehensive Human-Robot Interaction (HRI) studies illustrate the benefits and values of the proposed rear-projected robotic platform over a virtual-agent with the same animation displayed on a 2D computer screen. The results indicate that ExpressionBot is well accepted by users, with some advantages in expressing facial expressions more accurately and perceiving mutual eye gaze contact.

To improve social capabilities of the robot and create an expressive and empathic social agent (affect-aware) which is capable of interpreting users' emotional facial expressions, we developed a new Deep Neural Networks (DNN) architecture for Facial Expression Recognition (FER). The proposed DNN was initially trained on seven well-known publicly available databases, and obtained significantly better than, or comparable to, traditional convolutional neural networks or other state-of-the-art methods in both accuracy and learning time. Since the performance of the automated FER system highly depends on its training data, and the eventual goal of the proposed robotic platform is to interact with users in an uncontrolled environment, a database of facial expressions in the wild (called *AffectNet*) was created by querying emotion-related keywords from different search engines. AffectNet contains more than 1M images with faces and 440,000 manually annotated images with facial expressions, valence, and arousal. Two DNNs were trained on AffectNet to classify the facial expression images and predict the value of valence and arousal. Various evaluation metrics show that our deep neural network approaches trained on AffectNet can perform better than conventional machine learning methods and available off-the-shelf FER systems.

We then integrated this automated FER system into spoken dialog of our robotic platform to extend and enrich the capabilities of ExpressionBot beyond spoken dialog and create an affect-aware robotic agent that can measure and infer users' affect and cognition. Three social/interaction aspects (task engagement, being empathic, and likability of the robot) are measured in an experiment with the affect-aware robotic agent. The results indicate that users rated our affect-aware agent as empathic and likable as a robot in which user's affect is recognized by a human (WoZ).

In summary, this dissertation presents the development and HRI studies of a perceptive, and expressive, conversational, rear-projected, life-like robotic agent (aka ExpressionBot or Ryan) that models natural face-to-face communication between human and emapthic agent. The results of our in-depth human-robot-interaction studies show that this robotic agent can serve as a model for creating the next generation of empathic social robots.

Acknowledgements

It is my greatest pleasure to acknowledge my deepest gratitude and appreciation to my advisor Dr. Mohammad H. Mahoor for supporting me in this incredible field of research and his endless commitment. I am gratefully indebted to Dr. Mahoor for his very valuable comments on this thesis. I believe that it is always an excellent privilege to work under his supervision.

Additionally, I would like to express my sincerest appreciations to my oral defense committee Dr. Kimon Valavanis, Dr. Matthew Rutherford, Dr. Timothy Sweeny, and Dr. Kateri McRae for their support and assistance to improve this project. I truly appreciate their time and consideration. In addition, I would like to offer my special thanks to Dr. Ron Cole, President of Boulder Learning Inc. (BLI). Our research was not possible without collaboration with BLI in developing the lifelike avatar.

I am also very grateful to all of the people in the Computer Vision and Social Robotics Lab at DU who have worked with me during this research, especially Hojjat Abdollahi, Behzad Hasani, and David Chan.

I must express my very profound gratitude to my mother, father, and sister who live in my home country, Iran. I have not seen them for several years. I sincerely thank them for their endless love, support, and encouragement throughout this process and my life.

Last but not least, I would like to thank and dedicate this thesis to my lovely wife, Rozhin, who has been a constant source of support and encouragement during my Ph.D. Understanding me best as a Ph.D. student herself, Rozhin has been my best friend with endless love, support, and encouragement through this agonizing period of life. I am truly thankful for having her in my life.

This research was supported in part by National Science Foundation Grants IIS-1111568, CNS-1427872, and IIP-1449747.

Table of Contents

1 Introduction **1**

2 ExpressionBot Design **7**

2.1 ExpressionBot: Mechanism and Design 10

2.1.1 Neck System 10

2.1.2 Display System 11

2.2 Animation 11

2.2.1 Lip Blending with Emotion 13

2.3 Calibration 15

2.4 New Design 17

2.4.1 Ryan 19

2.5 Conclusion 20

3 Human-Robot Interaction **21**

3.1 Initial Evaluation 22

3.2 Evaluation of Presence and Embodiment 25

3.3 Embodiment and Presence 29

3.3.1 Research Questions 31

3.4 Visual Speech 32

3.4.1 Related Work 33

3.4.2 Methodology 36

3.4.3 Visual Speech Experiment 36

3.4.4 Visual Speech Results 39

3.5 Facial Expressions 40

3.5.1 Related Work 41

3.5.2 Methodology 43

3.5.3 Facial Expressions Experiment 45

3.5.4 Facial Expression Results 46

3.6 Eye Gaze 49

3.6.1 Related Work 51

3.6.2 Methodology 53

3.6.3 Eye Gaze Experiment 56

3.6.4 Eye Gaze Results 58

3.7 Discussion and Conclusion 63
3.7.1 Conclusion 67

4 Affect Perception 70
4.1 Existing FER Systems 72
4.2 Facial Landmark Point Extraction 77
4.2.1 Active Appearance Model (AAM) 79
4.2.2 Bidirectional Warping for AAM Fitting 82
4.2.3 Experimental Results 86
4.2.4 Discussion and Conclusions 92
4.3 A New Deep Neural Network Architecture 93
4.3.1 Face Databases 98
4.3.2 Results 99
4.3.3 Discussion 103
4.4 AffectNet 104
4.4.1 Existing databases 107
4.4.2 Evaluation Metrics 113
4.4.3 Existing Algorithms 115
4.4.4 Facial Images from the Web 118
4.4.5 Annotation 119
4.4.6 Annotation Agreement 125
4.5 Baseline 127
4.5.1 Test, Validation, and Training Sets 127
4.5.2 Categorical Model Baseline 128
4.5.3 Dimensional Model (Valence and Arousal) Baseline 134
4.6 Conclusion 137

5 Affect-Aware Agent 140
5.1 Empathy 141
5.1.1 Empathy in Social Assistive Robots 142
5.2 Methodology 146
5.2.1 Automated FER System 149
5.2.2 Empathic Conversations 150
5.3 Social/Interaction Aspects Measurements 156
5.4 Results 157
5.5 Conclusion 160

6 Conclusion and Future Work 162

Bibliography 167

A Publications 192

B AffectNet **194**

C Affect-Aware Agent **199**

C.1 Empathy Stimuli . . . 199

C.2 Affect-Aware Questionnaire . . . 200

List of Tables

3.1 Confusion matrix of recognized expression on the physical agent 23
3.2 Confusion matrix of recognized expression on the screen-based agent . . . 23
3.3 Visual speech rating on the physical agent and the screen-based agent . . . 24
3.4 Summary and overview of literature comparing audio-visual speech in different conditions 34
3.5 Summary and overview of literature comparing perception of emotion in different conditions 42
3.6 Confusion matrix of the emotion recognition rate for different agent condition 47
3.7 Pairwise comparison of users' perception of facial expressions on different agent conditions 50
3.8 Summary and overview of literature comparing perception of eye gaze in different conditions 51
3.9 Average and proportional error with respect to human ground-truth for different agent conditions. 59
3.10 Pairwise comparison of users' perception of eye gaze at different head rotations 63
3.11 Summary of role of embodiment and presence in perception of different facial cues. 67

4.1 RMSE of fitting on the left and right poses. 91
4.2 Percentage of fitted on the left and right poses. 91
4.3 Proposed DNN architecture Configuration 98
4.4 Number of images per each expression in databases 99
4.5 Average Accuracy (%) for subject-independent 100
4.6 Average (%) confusion matrix for subject-independent 101
4.7 Average Accuracy (%) on cross database 102
4.8 Subject-independent comparison with AlexNet results (% accuracy) 103
4.9 Summary and Characteristics of Databases in Affect Recognition 112
4.10 State-of-the-art algorithms on the databases listed in Table 4.9 117
4.11 Number of Annotated Images in Each Category 121
4.12 Percentage of Annotated Categories for Queried Emotion Terms (%) 122
4.13 Annotators' Agreement in Dimensional Model of Affect 125
4.14 Agreement Between Two Annotators in Categorical Model of Affect (%) . 125

4.15 F1-Scores of four different approaches of training AlexNet 132
4.16 Evaluation Metrics and Comparison of CNN baselines on Categorical Model 132
4.17 Confusion Matrix of Weighted-Loss Approach on the Test Set 133
4.18 Baselines' Performances of Predicting Valence and Arousal on Test Set . . 136

5.1 Confusion matrix of automated facial expression recognition 158
5.2 Subjects' facial expression during watching videos 158
5.3 Pairwise comparison of users' rating for different empathic agents 160

B.1 Agreement percentage between two annotators in categorical model 194
B.2 Samples of annotated categories for queried emotion terms 195
B.3 Samples of annotated images by two annotators 196
B.4 Number of annotated images in each range/area of valence and arousal . . . 198
B.5 Evaluation metrics and comparison of CNN baselines on the validation set . 198
B.6 Baselines' performances of predicting valence and arousal on the validation set 198

C.1 Affect-Aware Questionnaire 200

List of Figures

2.1 ExpressionBot's design and configuration. 12
2.2 Examples of some visemes and expressions 14
2.3 Calibration process. 16
2.4 Result of calibratio 17
2.5 New ExpressionBot's design and configuration 18
2.6 Ryan 19

3.1 Experimental setup and placement of subjects. 26
3.2 Presence and embodiment dimensions across three categories of experimental stimuli in the literature 28
3.3 The average accuracy of audio-visual speech perception in different conditions. 40
3.4 Six basic facial expressions at their maximum intensity 44
3.5 Different intensity level of surprised emotion on the virtual agent 45
3.6 The average accuracy of emotion perception in different conditions. 48
3.7 Schema and the variables used in the calculating eye gaze angle 55
3.8 Mask with flat eye region (left) and with angled eye region (right) 56
3.9 Perception of eye gaze room setup. 57
3.10 Eye gaze different conditions 59
3.11 Average error in perception of eye gaze for different head rotation and gaze shift 60
3.12 Estimated marginal means of gaze perception error for different head rotation angles and gaze target points 69

4.1 The Bidirectional Warping Algorithm 84
4.2 Some sample images of frontal, left and right poses from Multi-PIE dataset 87
4.3 Initial and fitted shapes of a sample image. 88
4.4 RMSE of fitting for variation of Projecting Out. 89
4.5 RMSE of fitting for variation of SIC. 89
4.6 Percentage of fitted images for variation of PO. 90
4.7 Percentage of fitted images for variation of SIC. 90
4.8 Network Architecture 96

4.9 Sample of the face registration taken from MultiPIE, SFEW, MMI, CK+ and DISFA 97
4.10 Training loss and classification accuracy on validation set 102
4.11 Sample images in Valence Arousal circumplex 106
4.12 Screen-shot of the software application to annotate AffectNet 118
4.13 Samples of queried images from the web and their annotated tags 122
4.14 Histogram of valence and arousal annotations 124
4.15 Samples of miss-classified images 135
4.16 Euclidean error of training valence and arousal. 136
4.17 RMSE of predicted valence and arousal using AlexNet and Euclidean loss . 137

5.1 Two different perspectives of studying empathy in social agents 144
5.2 Evaluation of empathic agent room setup 148
5.3 Conversation graph of the experiment 149
5.4 Building block of Residual learning in ResNet 150
5.5 Empathic conversation map after showing a video intended to elicit a happy emotion 152
5.6 Empathic conversation map after showing a video intended to elicit a surprise emotion 153
5.7 Empathic conversation map after showing a video intended to elicit a sad emotion 154
5.8 Empathic conversation map after showing a video intended to elicit a disgust emotion 155
5.9 The average user's rating and standard error of different empathic agents. . 159

B.1 Sample images in valence arousal circumplex with their corresponding values 197

Chapter 1

Introduction

Social intelligent robotics is a rapidly emerging field aiming to design robots that are able to communicate and interact with humans in a socially acceptable way (Breazeal, 2005; Dautenhahn, 2007). They often to achieve positive outcomes in diverse applications such as education, health-care, quality of life, entertainment, communication, and tasks requiring collaborative teamwork (Breazeal et al., 2016). These robots are becoming an integrated part of our daily lives. The population of robotic agents including social and humanoid robots made in 2008 was about 8.6 million units (Guizzo, 2010) with a projected annual growth rate of 17% (IDC, 2016). Despite significant progress towards the development of realistic social robotic agents, a number of problems remain to be solved.

First, current social robots either lack enough ability to have deep social interaction with human or they are very expensive to build and maintain. Social robots such as Paro (2004) have the robustness and cost-effectiveness for large-scale production, but lack the sophistication for deep social interaction. Other robots such as Kismet (2002) exhibit facial expressions and head and ear movement using mechanical components, however, once these mechanical platforms are built, they are fixed and cannot be readily modified. On the other hand, more human-like robots such as Geminoid (2011) possess profound capabilities for social interaction, but due to a large number of actuators in their mechatronic faces, they are expensive and maintenance-intensive for large-scale trials. Another potential problem in the design of robotic heads is the "Uncanny Valley" effect (Mori et al., 2012), where the effect of aesthetic design of a robot may influence the user's experience, perception, and acceptance of the robot.

Second, current social robots have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. Existing robotic platforms lack the capability of being perceptive and expressive as well as supporting natural spoken dialog. They often do not endow affect perception and most of the studies carried out with social robots are either done in a Wizard-of-oZ (WoZ) manner (Vardoulakis et al., 2012), or were limited to a specific scenario (Pineau et al., 2003).

To address these problems, we designed, manufactured, and evaluated of an affect-aware rear-projection robotic head, called ExpressionBot, with the capability of showing facial expressions, visual speech, eye gaze, and perception of user's facial expression, with the goal of closely modeling the dynamics of natural face-to-face communication. Our eventual goal is to provide the research community with a low-cost portable facial display hardware equipped with a software toolkit that can be used to conduct research leading to a new generation of robotic heads that model the dynamics of face-to-face communication with individuals in different social, learning, and therapeutic contexts. The proposed robotic head can then be integrated with a torso, arms, and even legs but the focus of this dissertation is the head and face design. To achieve this goal, three research/work streams is performed in this dissertation as:

1- Design, Implementation, and Study of a Rear-Projected Robotic Agent: Given the tremendous effort required to develop robot heads and the number of design choices that must be made, including aesthetic face design, mechanical design, and construction, hardware implementation, etc., it is often difficult to redesign or rebuild the head based on user experiences and limitations discovered during research. In addition, major obstacles in developing realistic robots faces lie with the actuators and the skin. The FACS system codes for approximately 40 primary facial muscles movements (AUs) that are involved in showing facial expressions and mouth movement during speech. These actions can be very subtle and quick, and many times mechanical actuators fail to mimic them. Also, due to cost and space constraints, android robotic heads have few actuators and their faces are relatively larger than an average head. Besides, the skin of the robot, which is often made of latex, makes unnatural wrinkles and folds on the robot face.

An alternative approach that overcomes many of these problems is to use state-of-the-art character animation technologies to create an animated human face (aka avatar) that can produce natural speech and facial expressions, and then project this avatar onto a face-shaped translucent mask. The mask and the projector can then be rigged onto a neck mechanism that can move like a human head. In this dissertation, we described the design and creation of a low-cost emotive robotic head, called ExpressionBot, for natural face-to-face communication. ExpressionBot consists of a simple neck system and a projector that projects a facial animation on a 3D translucent facial mask. By virtue of the computer graphics used to generate the avatar, highly realistic, accurate, and dynamic animations can be generated. These avatars can range from cartoon-like to photo-realistic faces and are usually able to show natural visual speech and facial expressions. The proposed robotic system, relative to mechatronic and android faces, is thus a highly flexible research tool, mechanically simple, and low-cost to design, build and maintain (the cost of the hardware system is about $1500).

Since an animation face is projected onto a mask, a big question is: *"What are the value propositions of a rear-projected robot compared with an on-screen animation?"*. In this dissertation, the above question is answered by studying different elements of Human-Robot vs. Human-Animation Interaction between the rear-projected robotic head and a virtual-agent with the same animation displayed on a 2D computer screen. At first, individuals' experiences of interpreting the facial expressions and the proposed visual speech of ExpressionBot is compared with the facial animation on the computer screen. We then distinguished the role of the robot's embodiment from its physical presence in perception of three facial cues (i.e., visual speech, facial expressions and eye gaze). In particular, three different conditions (i.e., copresent of the robot, telepresent of the robot, and virtual agent) were studied to answer whether the embodiment of the robot has any interaction value proposition compared with an on-screen animation.

2- Developing a New Affect Perception System: Although SARs are finding their place in our society as artificial pets, entertainers, and tools for therapists, current technologies have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. To achieve this potential, research must imbue robots with the emotional and social capabilities—both verbal and non-verbal—necessary for rich and robust interaction with human beings. Facial expression, which plays a vital role in social interaction, is one of the most important nonverbal channels through which Human-Machine Interaction (HMI) systems can recognize humans' internal emotions. Due to the importance of facial expression in designing HMI and Human-Robot Interaction (HRI) systems, numerous computer vision and machine learning algorithms have been proposed for automated Facial Expression Recognition (FER). The majority of the techniques for automated affective computing and FER are based on supervised machine learning methodologies which require annotated image samples for training.

Recently, due to an increase in the availability of computational power and increasingly large training databases to work with, the machine learning technique of neural networks has seen a resurgence in popularity. Deep Neural Networks (DNN), also called deep learning, obtained state of the art results in several fields of computer visions. Given DNN's performance, we proposed a new DNN architecture for facial expression recognition. The proposed DNN is trained on seven well-known publicly available databases, and obtained significantly better than, or comparable to, traditional convolutional neural networks or other state-of-the-art methods in both accuracy and learning time.

The databases used to train the proposed DNN FER system mainly contain posed expressions acquired in a controlled lab environment. Hence, the proposed DNN FER system lack enough generality when used in uncontrolled HRI system. Since eventual goal of the proposed robotic platform is to interact with users in an uncontrolled setting (aka *"in the wild"* setting), where there is a high variation in scene lighting, camera view, image resolution, background, subjects head-pose and ethnicity, we created a database of facial *Affect* from the Inter*Net* (called **AffectNet**) by querying more than one million images from different search engines using emotion-related tags in six different languages. We then proposed a DNN baseline to classify the facial expression images and predict the value of valence and arousal. Various evaluation metrics show that our deep neural network baselines can perform better than conventional machine learning methods and off-the-shelf facial expression recognition systems.

2- Creating an Affect-Aware Robot: Finally, we integrated our automated FER system into the spoken dialog system of our robotic platform to extend and enrich the capabilities of ExpressionBot beyond spoken dialog and create an affect-aware robotic agent that can measure and infer users' affect and cognition. We evaluated whether this integration can improve social/interaction aspects of our agent with users. We designed a series of HRI experiments, in which the subjects watched some videos to evoke their emotions and the robot asked them to describe each video in one word. During watching the videos, the robot recognized subjects' facial expressions and engaged them in conversation based on the perceived facial expressions. We then measured the accuracy of the automated FER system on the robot when interacting with different human subjects as well as three social/interaction aspects, namely task engagement, being empathic, and likability of the robot.

The remainder of this dissertation is organized as follows: Chapter 2, overviews some existing social robot platforms, and then describes the mechanism and design of the proposed robotic platform. Chapter 3 presents a deep HRI study using the developed rear-projected robotic agent. An initial user evaluation test is studied to evaluate individuals' experiences and impressions of the ExpressionBot. Then, three major elements of face-to-face communication (i.e., visual speech, facial expressions and eye gaze) are studied in three different conditions (i.e., copresent of the robot, telepresent of the robot, and virtual agent) to evaluate the role of embodiment and presence of the robot in comparison with the same animation projected on a 2D screen.

Chapter 4 explains the proposed DNN architecture and the process of creating and annotating *AffectNet*. Chapter 5 discusses the process and evaluation of integrating the proposed automated FER system into spoken dialog of the developed robotic platform to create an affect-aware robotic agent. Finally, Chapter 6 concludes the finding of this research and discusses future paths of this investigation.

Chapter 2

ExpressionBot Design

There are many designs for robotic faces ranging from 2D graphical avatars to mechanically controlled robotic faces. These designs fit into four main categories:

1. Mechatronic Faces
2. Android Faces
3. Onscreen Avatars
4. Light-Projected physical Avatars

Mechatronic robotic faces are physically implemented robots that use mechatronic devices and electric actuators to control facial elements. Kismet (2002) is one the first and famous expressive mechatronic robots with many features such as eye lids, eye brows, lips and even expressive ears. Another example is the Philips iCat (van Breemen et al., 2005) which has a cat-like head and torso with mechanical lips, eye lids, and eye brows. Mechatronic robotic faces have the advantage of being 3D, but they are inflexible, unrealistic, and have limited ability to display facial expressions and speech. These faces look very much like a stereotypical robot rather than a human face.

Android faces are other physically implemented robots that are originated from Animatronics. They have a larger number of mechatronic actuators controlling a flexible elastic skin; therefore they look more realistic and seem more like a human rather than a robot. Example of android faces are Albert-Hubo (2005), HRP-4C (2009), Geminoid (2011), and Zeno (2009). It is an interesting research question as to whether android faces that closely model human looks and behaviors will enter the uncanny valley as their realism mimics humans. Due to larger number of actuators and their interaction with skin, they look more expressive than mechatronic faces. However, they are mechanically very complex, expensive to design, build and maintain.

The on screen avatar class, such as Grace (Gockley et al., 2004) and Second-Life (2003) are the simplest and earliest robotic faces. Animations for these models can be made by developing a model for each expression, morphing between them, and then rendering the result to a computer screen. Despite their low cost and high flexibility, they naturally have several limitations due to using a flat display as an alternative to a three dimensional physical head. For example, aside aesthetic unpleasantness, they suffer from lack of establishing mutual gaze (due to Mona Lisa effect) and physical embodiment that both play vital roles in natural and realistic face-to-face communication.

The final category, and the focus of our research, is the light-projected physical avatars; these consist of translucent 3D masks with 2D/3D avatars projected onto them. The avatar can be projected from the front of a facial mask (forward-projection) or back of a translucent mask (rear-projection). Since an animation is projected onto a mask, the robotic face can range from cartoon-like to photorealistic. Light-projected physical avatars are thus a highly flexible research tool, relative to mechatronic and android faces. Factors such as engagement, embodiment, believability, credibility and realism can be investigated based on the appearance and behaviors of the 3D animated models, and the robotic head and neck movement. Moreover, such a system can avoid the Mona Lisa effect (Todorović, 2006) and hence users can correctly perceive the robot's eye gaze direction. Additional features of robotic avatars include relatively low development cost, low power consumption, potentially low weight and fast reaction.

One of the early examples of rear-projection physical avatars is the Dome robot (Hashimoto and Morooka, 2005) where a cartoonish animated face is projected onto a dome-shaped mask. The dome mask makes the image and display calibration process easy. However, it lacks human face realism and the results appear cartoonish. The Lighthead robotic face (Delaunay et al., 2011) is another example that also projects an animation onto a face-shaped translucent mask, resulting in a more realistic appearance. It is capable of displaying a wide range of facial expressions and emotions. Kuratate et al. (2011) presented a mask head robot, called Mask-bot, that generates visual speech, head movements, facial expressions and eye movements. Then later, Pierce et al. (2012) introduced Mask-bot 2i with an automatic approach for projection calibration by using a series of gray-coded patterns in a calibration booth which supports interchangeable masks. Both Mask-bot and Mask-bot 2i use human talking head animation that is photo realistic, which is not as flexible as computer animation.

Al Moubayed et al. (2012) introduced Furhat robot that utilizes computer animation to deliver facial movements on a 3D translucent facial mask. They also studied the perception of animation's gaze on 3D projected against flat screens and demonstrated the limitations of flat screens in delivering accurate direction of gaze due to Mona Lisa effect, which limits having situated, multiparty interaction in onscreen Avatars. Furhat uses a pan-tilt unit for the neck system which has only 2 DOF pitch and yaw. Also Furhat uses a mirror instead of a fish eye lens which makes the robot to have a larger form factor.

Another approach of creating light-projected avatars is to use forward-projection instead of rear-projection. Lincoln et al. (2009) introduced a forward-projection animation system called Shader Lamps Avatar where the dynamic motion of a real person is captured with a camera and motion sensors; the human motion is then used to control a humanoid animatronic which is projected on a mask by a front projector. Front projections robotic avatars are able to portray fully side view of the character, however the whole system is much larger than the rear-projection robotic avatars.

The remainder of this Chapter is organized as follows: Section 2.1 reviews the mechanism and design of the neck system of the ExpressionBot. The animation system and the process of lip blending with facial emotion is elaborated in Sec. 2.2. Section 2.3 reviews the proposed calibration method to calibrate the animation on the robot's facial mask. A new design of ExpressionBot that overcome some of the flaws in previous proposed version in introduced in Sec. 2.4, and Sec. 2.5 conclude this Chapter.

2.1 ExpressionBot: Mechanism and Design

The ExpressionBot consists of three main sections, the neck control system, the display system and the animation application. The neck system controls the projector and mask position allowing it to be rotated by the application to track faces and head gestures. The display system consists of a small projector with a fish eye lens that projects the animation on a human like (head shaped) face mask. The animation application displays a face animation along with speech and emotion to be projected on to the mask.

2.1.1 Neck System

The neck mechanism of our existing prototype has three degrees of freedom (DoF) providing a total of 150° of yaw (x-y plane), 30° of pitch (x-z plane) and 30° of roll (y-z plane). Solidworks CAD tool was used to design and maximize the range of motion, resulting in a light, compact, and quiet mechanism. These design constraints were achieved using a 6"×6" footprint, low friction plastic gears and brushless servomotors. Also, the small footprint allows the neck and projector to be easily shrouded by the mask, and allows the user to control the distance from the mask to the lens in order to project the clearest possible image.

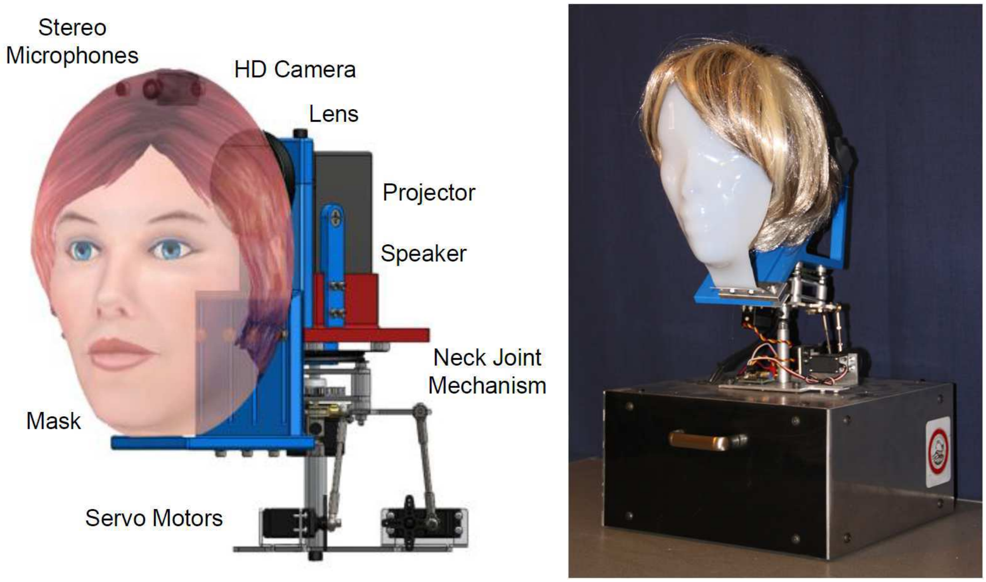
2.1.2 Display System

Our system uses a Dell DLP M110 portable projector. The projector is capable of up to 300 ANSI Lumens under normal indoor illumination conditions, can display a maximum resolution of 1200×800, and has a 10000:1 contrast ratio. Attached to the projector is a Nikon Fisheye Converter FC-E8 which provides a viewing angle of approximately 183 degrees. This allows the projector to display to the whole mask from a relatively close distance.

To create the mask we designed a mold using the 3D model of the neutral face in Autodesk Maya. We 3D printed this mold and used it to vacuum form a 1/8 inch sheet of white translucent acrylic plastic. Then, we added a metal band from top of the mask to the projector, which allows us to mount a wig on the robot's head. This makes the ExpressionBot more aesthetically pleasant and natural, and covers the lights coming out from the sides of the mask due to fish eye lens wide projection-angle (see Fig. 2.1).

2.2 Animation

We developed a face animation in C# .Net for accurate natural visual speech and show expression based on multi-target morphing method (Ma and Cole, 2004). Recorded utterances are processed by the Bavieca speech recognizer (Bolanos, 2012), which receives the sequence of words and the speech waveform as input, and provides a time-aligned phonetic transcription of the spoken utterance. The aligned phonemes are represented using the International Phonetic Alphabet (IPA), a standard that is used to provide a unique symbolic



(a) (b)

Figure 2.1: ExpressionBot's design and configuration.

notational for the realization of phonemes in all of the world's languages (Association, 1999). As IPA is intended as a standard for the phonemic and phonetic representation of all spoken languages, having IPA in our system will allow us to add other languages easily as long as the speech recognizer is trained for that language.

For a given language, visually similar phonemes are grouped into units called visemes. For example the consonants /b/, /p/ and /m/ in the words "buy," "pie," and "my" form a single viseme class. We categorized English phonemes into 20 viseme classes. These classes represent the articulation targets that lips and tongue move to during speech production. A graphic artist designed 3D models of these viseme classes in Maya. Figure 2.2, demonstrates some visemes used in our animation system. Finally, natural visual speech is obtained by blending the proper models corresponding to each part of speech with different weights.

To achieve a smooth and realistic look, we used a kernel smoothing technique. During speech production, the avatar system receives the time-aligned phonetic input from Bavieca system, converts the phonetic symbols into the corresponding visemes, which specifies the movements of the mouth and tongue, synchronized with the recorded or synthesized speech. The algorithm models coarticulation by smoothing across adjacent phonemes. We use the Epanechnikov kernel (Epanechnikov, 1969) to pull the weights for each viseme associated with the current time value and set the weights for those visemes' morph targets.

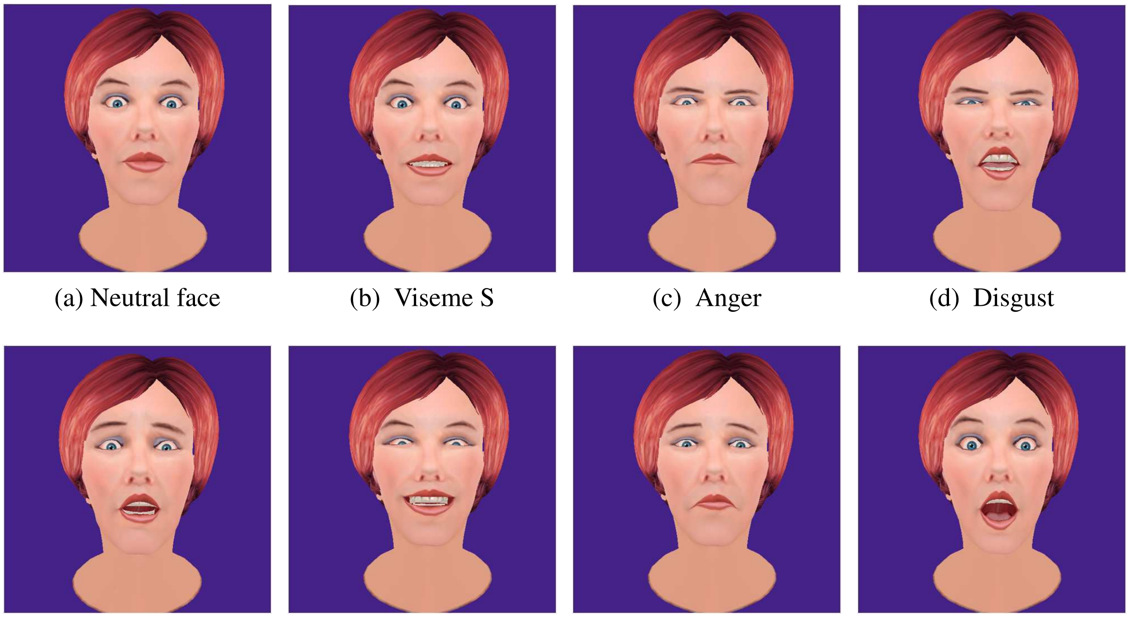
Using the kernel technique resulted in smoother and more natural looking animations; however, when utterances included the labial phonemes /b/, /m/, /p/, which are accompanied by lip closure, the smoothing algorithm prevented the lips from closing when the duration of the labial phoneme is very short (e.g., 5 msec.) and the adjacent phoneme targets caused the lips to be open (e.g., /ɒ/ as in "mama"). To force lip closure for the labials, we extended the duration of labial visemes to include the closure interval (the period of relative silence before the sound is released, thus increasing the chance that at least one frame consisting of just the labial viseme will appear.

We designed the models in three portions: eyes, face and hair. This design allows them to be interchangeable and customizable, and gives us the ability to design any number of characters to easily change the robot's appearance. The system has the ability to control eye gaze independently of the visual speech and facial expression animation, and thus enables functionality to control eye gaze (e.g., in concert with face tracking).

2.2.1 Lip Blending with Emotion

In order to blend the expressions with the lip movement, the animation uses the following formula to generate facial expressions based on the current viseme and emotion morph targets:

$$F_j = F_c + \lambda_j (F_j^{max} - F_0) \tag{2.1}$$



(a) Neutral face (b) Viseme S (c) Anger (d) Disgust

(e) Fear (f) Joy (g) Sadness (h) Surprise

Figure 2.2: Examples of some visemes and expressions

where F_c represents the current viseme, F_j^{max} is the desired expression model at the maximum intensity, F_0 is the Neutral model. The parameter $\lambda_j \in [0, 1]$ is the intensity of the j^{th} expression model F_j. The graphic artist designed 3D models of six basic expressions (i.e., anger, disgust, fear, joy, sadness and surprise) in Maya based on Facial Action Coding System (FACS) (Ekman and Friesen, 1977). For example joy involves Cheek Raiser (AU 6) and Lip Corner Puller (AU 12) and sadness involves Inner Brow Raiser (AU 1), Brow Lowerer (AU 4) and Lip Corner Depressor (AU 15).

In order to blend the expressions with the lip movement, adding weight to the emotion morph targets without regard to the movements of the face caused by speech production will result in unnatural looking facial expressions. For example, combining the surprise expression which causes the mouth to be fully open, conflicts with the production of phonemes like /b/, /f/ and /v/ that are produced with the lips closed or nearly closed. Combining the joy emotion with puckered mouth visemes such as /o/ will also result in visual speech and expressions that are not natural and are perceived as abnormal or creepy. To overcome this problem, we designed a table that provides a viseme weight factor and a maximum emotion weight for every viseme emotion combination. These values are adjusted empirically for each combination.

We separated the facial expression morph targets into upper and lower face morph targets; the upper face includes everything from the tip of the nose upwards. The lower face includes the region below the nose; mainly the lips, lower cheeks and chin. This partitioning of the face enables us to adjust the weight of just the lower face morph target weights so that the upper face remains consistent with the morph targets of desired expressions. In addition, for labial and labiodental visemes (those for the letters m, b, p, f and v) that require the avatar's lips to be closed or nearly closed to look natural, we developed visemes pre-blended with the open mouthed emotions. These are used to replace the viseme and lower face expression when they come up in combination.

2.3 Calibration

Due to the projector and fish eye lens distortions, the resulting direct projection of the animated face model on the mask appears distorted (See Fig. 2.4). Hence, we need to rectify the projection so the image appears undistorted and the facial regions (e.g., eyes, mouth) of the facial model are projected to the desired position on the mask. In order to achieve a smooth animation displaying at 30 fps, we decide to distort the original Maya models rather than rectifying the projection at run time in each frame.

Assuming N is the neutral model in model coordinates (Fig. 2.3.a), $S = N \times WVP$ is the displayed projection in the screen coordinates where WVP matrix is the multiplication of the world, view and projection matrices, respectively (Fig. 2.3.b). Assuming $M = project(S)$ is the projected model on the mask (Fig. 2.3.c), we aim to find N' (the distorted

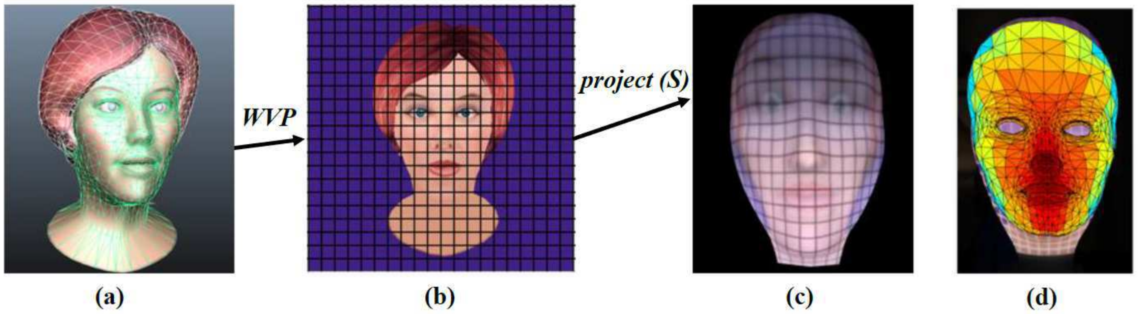
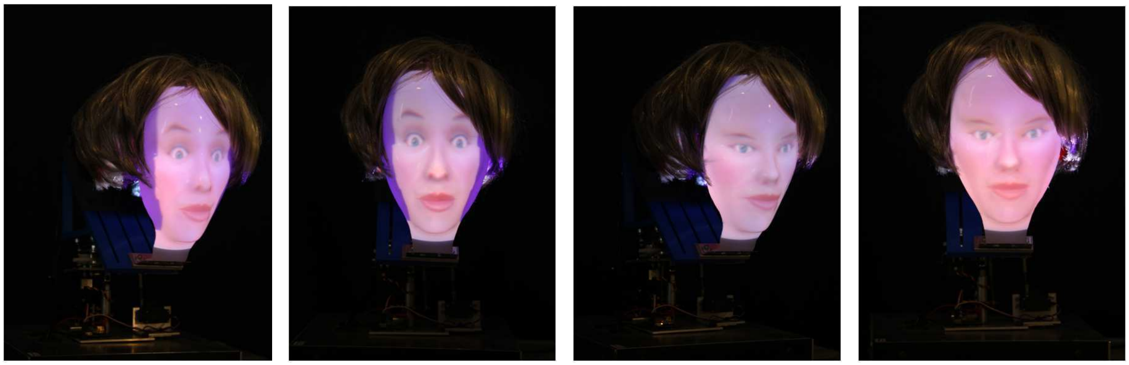


Figure 2.3: Calibration process.

neutral model) such that $M' = project(S')$ looks undistorted on the mask, where $S' = N'WVP$. In order to estimate the function $project(.)$, we create a checkerboard in the screen coordinates (Fig. 2.3.b) and projected on the mask (Fig. 2.3.c). Then, we define a piecewise homography mapping between the corresponding rectangles of the mask and the triangles displayed on the screen. To find the undistorted neutral model on the mask, M', we apply an affine transformation to place the mold model, used to create the mask in the vacuum machine, on an image of the mask (Fig. 2.3.d). Afterwards, we apply the piecewise homography on the mold model and replace the corresponding vertices of the neutral model in the screen coordinates, S, with distorted mold model to estimate S'. We finally use $N' = S'WVP^{-1}$ as the neutral model in our application. Figure 2.4 shows the results of our calibration.

In order to improve the rendering speed and overcome the limitations of the rendering software libraries, we develop a blend-shape system that encodes the vertex position differences between each visemes/expression and the neutral model. The application renders the neutral model with these position differences rather than just blending the models. Because of this fact, it is unnecessary to apply our calibration procedure to all the viseme and expression models.



(a) (b) (c) (d)

Figure 2.4: Result of calibration: (a) and (b) show projection without calibration, (c) and (d) show projection with calibration from side and frontal view.

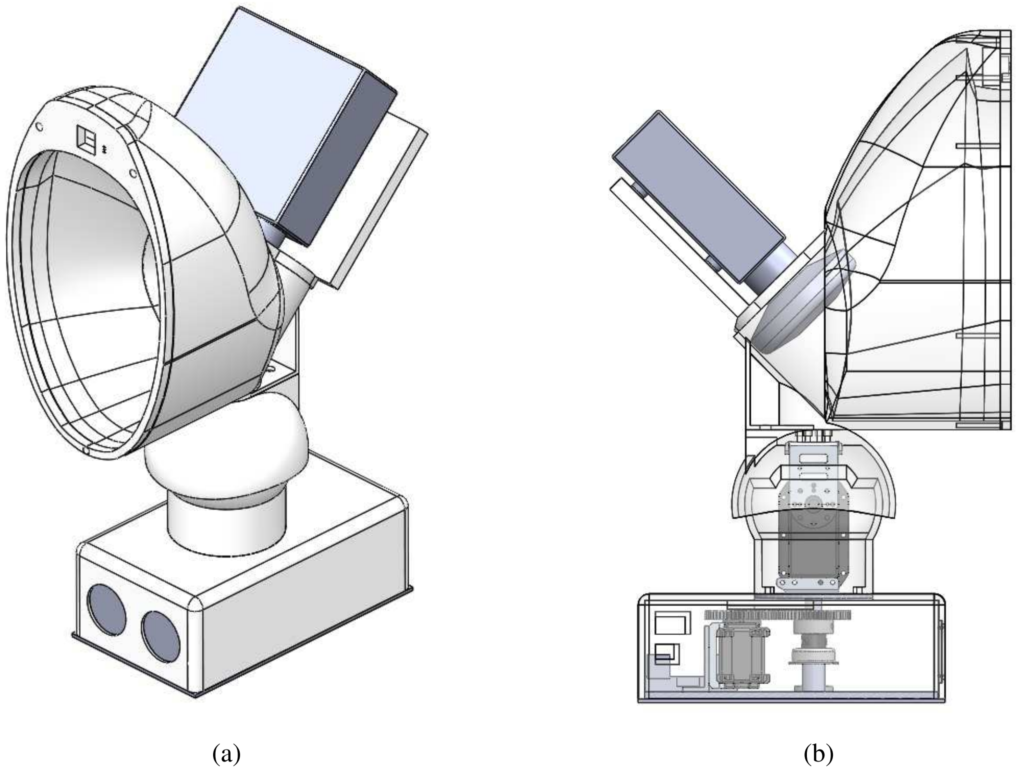
2.4 New Design

The design of ExpressionBot discussed in Sec. 2.1.1 and 2.1.1 had the following flaws:

1. **Neck System:** The neck mechanism of the first prototype had three degrees of freedom (DoF). However, three degrees of rotations were not fully independent, i.e., yaw (x-y plane), pitch (x-z plane) controlling servos depended on the position of the roll (y-z plane) servo motor. This made programming the servos difficult. Also, the neck system was not compact and made the system unnaturally large.

2. **Lip Protrusion:** Although the mask was built based on the 3D model of the neutral face, the lips protrusion on the mask made the visual speech unnatural since the mask is static, the jaw and lip movements are only optical and are not moving according to the speech phonemes movement.

3. **Head enclosure:** Since the fish-eye lens had a viewing angle of approximately 183 degrees, the light were scattered from the sides. Therefore, we mounted a wig to hide the bar or band extending from a top of the face mask to the projector and cover any stray light coming from the sides of the face mask (See Fig. 2.1). Since, the proposed



(a) (b)

Figure 2.5: New ExpressionBot's design and configuration. [Patent: Mahoor et al. (2015)]

robotic system is able to project any animation (e.g., male, female, cartoonish, etc.), the wig might not fit different animation character.

To overcome these problems, and make the system more compact, we re-design the ExpressionBot. The new design contains a head enclosure to prevent any projected light from being scattered from the sides of the robotic head, which may make the projected facial image on the mask look brighter. Also, the neck system is replaced with a pan-tilt unit coupled with the neck system and configured to move the head enclosure and the face mask. This makes the whole system more compact and also affordable. In addition, the lips protrusion were removed by smoothing out the lips of the neutral face 3D model. Figure 2.5 shows the solid-work model of the new system. As it is shown, the model is more compact compared with previous version shown in Fig. 2.1.

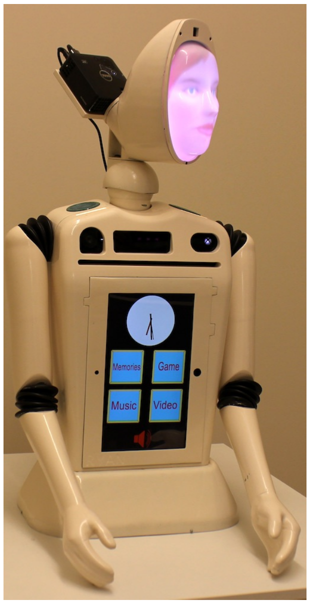


Figure 2.6: Ryan

2.4.1 Ryan

Based on the new design, and through a license agreement with the University of Denver, a new robotic platform is developed at DreamFace-Tech. (2015) called Ryan. Ryan (shown in Fig. 2.6) has a torso equipped with a 10” LCD touch screen which can be used to gather sensory input, display videos, and play games with users. Ryan is equipped with a Microsoft Kinect to track users’ movements and two stationary arms for an increased sense of realism. The neck has two degrees of freedom (DoF) providing a total of 180° of yaw, and 45° of pitch. The neck system controls the projector and mask position allowing it to be rotated by the robot application to track faces and head gestures. The same animations presented in Section 2.2 and calibration algorithm presented in Section 2.3 is used to calibrate the animation on the mask and rectify the distortion of the projector and the fish-eye lens [1]. We used Ryan in HRI studies (Chapters 3 and 5) as it is more aesthetic and has torso.

[1] Ryan platform and its new features are developed at DreamFace tech LLC. and it is not part of this dissertation.

2.5 Conclusion

Major obstacles in developing realistic robots faces lie with the actuators and the skin. The FACS system codes for approximately 40 primary facial muscles movements (AUs) that are involved in showing facial expressions and mouth movement during speech. These actions can be very subtle and quick, and many times mechanical actuators fail to mimic them. Also, due to cost and space constraints, android robotic heads have few actuators and their faces are relatively larger than an average head. Besides, the skin of the robot, which is often made of latex, makes unnatural wrinkles and folds on the robot face.

In this Chapter, we described the design and creation of a low-cost emotive robotic head, called ExpressionBot, for natural face-to-face communication. ExpressionBot consists of a simple neck system and a projector that projects a facial animation on a 3D translucent facial mask. Hence, the rear-projection robotic platform can portray natural and realistic facial movement, as advanced computer graphic system can easily animate them. Since an animation is projected onto a mask, the robotic face can range from cartoon-like to photorealistic. The proposed robotic system, relative to mechatronic and android faces, is thus a highly flexible research tool, mechanically simple, and low-cost to design, build and maintain (the cost of the hardware system is about $1500).

The developed robotic head represents a new level of integration of emotive capabilities that enables researchers to study socially emotive robots/agents that can generate spoken-language, show emotions, and communicate effectively with people in a natural way as humans do. Such systems can be applied in many domains including health-care, education, entertainment, and home-care. It will also be an ideal platform for designing a new generation of more immersive and effective intelligent tutoring and therapy systems, and robot-assisted therapeutic treatments.

Chapter 3

Human-Robot Interaction

In Chapter 2, the mechanism and design of the proposed robotic platform are discussed. As shown, the proposed robotic is a mechanically simple and low-cost platform that is built on the rear projection of an animation on a 3D translucent facial mask. It is known that, the perception of 3D objects that are displayed on 2D surfaces is influenced by the Mona Lisa effect (Todorović, 2006). In addition, physical embodiment can make a difference in perception of social robots (Dautenhahn et al., 2002; Wainer et al., 2006), which lacks in on screen avatar class. However, a big question is remained unanswered, as:

"What are the value propositions of a rear-projected robot compared with an on-screen animation?"

In this Chapter, the above question is answered by studying different elements of Human-Robot vs. Human-Animation Interaction. At first, individuals' experiences of interpreting the facial expressions and the proposed visual speech of ExpressionBot is compared with the facial animation on the computer screen. During these experiments, the users were in front of the robot, and it was not clear whether the users benefited from the physicality of the robot or they were under the impression of its physical presence. We then distinguished the role of the robot's embodiment from its physical presence in three major facial cues (i.e., visual speech, facial expressions and eye gaze). In particular, three different conditions (i.e., copresent of the robot, telepresent of the robot, and virtual agent) were studied to answer whether the embodiment of the robot has any interaction value proposition compared with an on-screen animation.

3.1 Initial Evaluation

In order to evaluate individuals' experiences and impressions of the ExpressionBot, we designed and conducted three experiments. Participants were 23 typical adults, 9 female and 14 males, with age range 18-51 years (Mean= 27.26, SD=7.79) and a variety of ethnicities (19 Caucasian, 2 Asian, 2 Hispanic).

Hereafter, we refer to the 3D computer character on the computer screen as the screen-based agent, and the projection of the 3D model onto the robotic head as the physical agent. We used a 23" LCD display to display the screen-based agent at the same size as the physical agent.

The objective of the first experiment was to assess how accurately subjects were able to interpret the projected facial expressions. Participants watched the robotic agent and the screen-based agent in two different sessions randomly (i.e. some participants observed the physical agent first while the others watched the screen-based agent first). A series of six basic emotions (joy, sadness, surprise, disgust, fear and anger) were displayed in random order. Each expression was displayed one time for about 5 seconds. The subject was then asked to select one of the six categories. They could also respond "none," if they were unable to assign the facial expression to one of the six categories.

Tables 3.1 and 3.2 present confusion matrices of the intended and classified expressions displayed on the physical agent and the screen-based agent, respectively. Comparing the percentages reported in these tables shows that the surprise and sad emotions were recognized perfectly (100% recognition rate) in both agents by the participants. The joy emotion was recognized perfectly when displayed on the physical agent but was recognized 92% of the time on the screen-based agent. Interestingly, Anger was recognized correctly 85% of the time for the physical agent, and only 38% of the time for the screen-based agent. Disgust was classified as anger more often than it was classified as disgust for both agents, and fear was recognized correctly over 50% of the time in both agents, and confused most often with sadness. In sum, the results showed high recognition rates for Joy, Sadness and Surprise in both agents, lower and similar recognition rates for Disgust and Fear in the two agents, and superior performance for Anger when displayed on the physical agent.

Table 3.1: Confusion matrix of recognized expression on the physical agent

%	Joy	Anger	Sadness	Disgust	Surprise	Fear	None
Joy	**100**	0	0	0	0	0	0
Anger	0	**85**	0	10	0	0	5
Sadness	0	0	**100**	0	0	0	0
Disgust	0	60	0	**40**	0	0	0
Surprise	0	0	0	0	**100**	0	0
Fear	0	0	35	10	0	**55**	0

Table 3.2: Confusion matrix of recognized expression on the screen-based agent

%	Joy	Anger	Sadness	Disgust	Surprise	Fear	None
Joy	**92**	0	0	8	0	0	0
Anger	0	**38**	8	46	0	8	5
Sadness	0	0	**100**	0	0	0	0
Disgust	8	46	0	**38**	8	0	0
Surprise	0	0	0	0	**100**	0	0
Fear	0	0	38	0	8	**54**	0

In the second experiment, we evaluated the proposed method for visual speech and examined subjects' judgments of speech production quality using the physical agent. Two short segments of speech were used in this experiment. Segment 1 was a seven second interval of a Margaret Thatcher's speech with length of 11 seconds while segment 2 was a seven second interval of Microsoft Anna synthetic speech. We chose these segments to cover a variety of length, speed, accent, and different phonemes (vowels, consonants and labial phonemes). Each speech was played two times with different lip synchronization approaches: 1) A basic approach where at each phoneme only the corresponding viseme was displayed without any kernel smoothing; 2) The proposed approach described in Section 2.2 where lip closure was enforced in labial phonemes and kernel smoothing was applied.

We asked the participants to rate how realistic the visual speech looked on a scale from 0 to 5, where 0 being unrealistic and 5 being very realistic. Table 3.3 shows the evaluation of the two speech segments for the two different visual speech approaches displayed on the physical agent and screen-based agent. One-tail paired T-test analyses were conducted, where the results show that there was not significant preferences between the physical and on-screen agents. However, the T-Test analysis indicated a significant preference for the proposed approach for synchronizing lips with speech over basic approach (p=.001 and p=.0002 on the physical agent and p=.0933 and p=.0067 on the screen-based agent for the speech segments 1 and 2, respectively).

Table 3.3: Average (STD) values of visual speech rating on the physical agent and the screen-based agent

	Physical Agent		**Screen-based Agent**	
	Basic	Proposed	Basic	Proposed
Speech 1	3.04 (0.80)	3.85 (0.85)	3.06 (0.79)	3.53 (0.91)
Speech 2	2.50 (0.88)	3.45 (0.75)	2.42 (0.93)	3.21 (0.69)

Inspired by the experimental setting in (Al Moubayed et al., 2012), we evaluated the perception of the eye gaze direction of the physical agent and screen-based agent. In this experiment, five subjects were simultaneously seated around an animated agent in two separate sessions; one session to examine the screen-based agent and another session for the physical agent (see Fig. 3.1). The seats were positioned at -45°, -25°, 0°, 25°, 45°, where 0° is the seat in front of the agent. The distance of the subjects to the agent was five feet.

In the first setting (called eye gaze only), the agent's head looked straight and only the eye gaze was shifted towards each subject. After each shift, all the subjects were asked to report the subject number that the agent was looking at. We repeated each experiment three times with 10 random eye gazes each time. In this setting, the subjects perceived the direction of the eye gaze 50% (SD=24%) of the times correctly for the screen-based agent and 88% (SD=13%) of the times correctly for the physical agent ($p<0.005$).

In the second setting (called eye gaze plus arbitrary head movement), the agent rotated its head and shifted its eye gaze randomly at the same time, but the head was not necessarily towards the subject of interest. Then, after each head movement and gaze shift, all the subjects reported the subject number that the agent was looking at. We repeated each experiment three times with 10 random eye gazes each time. In this setting, subjects perceived the direction of the eye gaze 43% (SD=18%) of the time correctly for the screen-based agent and 77% (SD=15%) of the times correctly for the physical agent ($p<.000017$). These results show that compared with the screen-based agent, the subjects perceived the eye gaze direction produced by the robotic face more accurately in both the "eye gaze only" and "eye gaze plus arbitrary head movement" settings.

3.2 Evaluation of Presence and Embodiment

Socially intelligent agents are becoming an integrated part of our daily lives. This is owing to advancements in computer technology, artificial intelligence, and recent innovations in virtual reality and computer graphics. The population of robotic agents including social and humanoid robots made in 2008 was about 8.6 million units (Guizzo, 2010) with a projected annual growth rate of 17% (IDC, 2016). Virtual agents, on the other hand, have

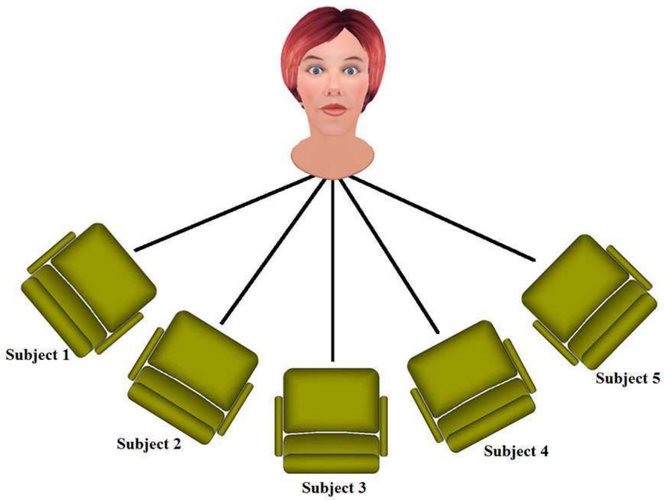


Figure 3.1: Experimental setup and placement of subjects.

received considerable attention in recent years as social agents (e.g. for museum guidance (Kopp et al., 2005), education (Vala et al., 2007), entertainment (Hartholt et al., 2009), and training for job interviews (Hoque et al., 2013)) due to the flexibility of computer rendered faces and the ubiquity of computer screens on mobile devices. Virtual agents are often used when a physical task or interaction such as moving objects is unnecessary. As robotic technologies are focusing more on improving social interaction with users, determining which kinds of robots or virtual agents are best suited for social interaction becomes increasingly important. One fundamental research question is what would be the difference between virtual agents and robots in terms of human interaction, particularly in perceiving major elements of face-to-face communication (both verbal and non-verbal facial cues and skills).

The most salient difference between a robot and a virtual agent on a computer screen is physical embodiment. Several investigations have compared various elements of social interaction among robots and virtual agents (Cassell, 2000; Ju and Sirkin, 2010; Kidd and Breazeal, 2004; Walker et al., 1994), and the majority of these investigations suggested that the physicality of the robot benefits user interaction. However, in the majority of these experiments, a robot with physical embodiment was physically present in front of the subjects. This is potentially problematic since the subject's percepts and evaluations may be affected not only by the robot's embodiment but also by its presence.

Some researchers evaluated the role of presence by comparing a robotic agent with its telepresence or an animated, computer-rendered of the robot (Li, 2015), however, few have compared all three conditions in a same experiment/platform (See Fig. 3.2). Also, the majority of these studies compared the influence of these agents on social elements such as likability (Kiesler et al., 2008), enjoyment (Wainer et al., 2007), etc. by requiring subjects to complete a questionnaire after interaction in the lab. Although the reliability of questionnaires can be validated by measurements such as Cronbach's Alpha (Cronbach, 1951), self-report may be an inaccurate quantitative measure, especially with small sample sizes. Hence, better quantitative measures are necessary to determine whether a physically present robotic agent can produce different, and perhaps superior experiences compared to a screen-based version of the same robot.

In the following sections, we studies the role of embodiment and presence in human perception of a rear-projected robot's facial cues. We used Ryan (the extended version of ExpressionBot explained in Chapter 2, Section 2.4.1), since the same animation from a virtual agent could be projected to this robotic face, thus allowing comparison of the virtual agent's animation behaviors with both telepresent and physically present robotic agents. Because face-to-face communication is an important method of social interaction which plays a major role in individuals' socialization and experience (Kendon et al., 1975), we focus on three major elements of face-to-face communication—visual speech, facial expression, and eye-gaze. We leverage three different agency conditions (copresent robot, telepresent robot, and virtual agent) to evaluate whether the embodiment and presence of a social robot provides any extra value for discriminating these social cues compared with an on-screen animation. Similar to other robotic platforms, Ryan has some limitations (e.g.,

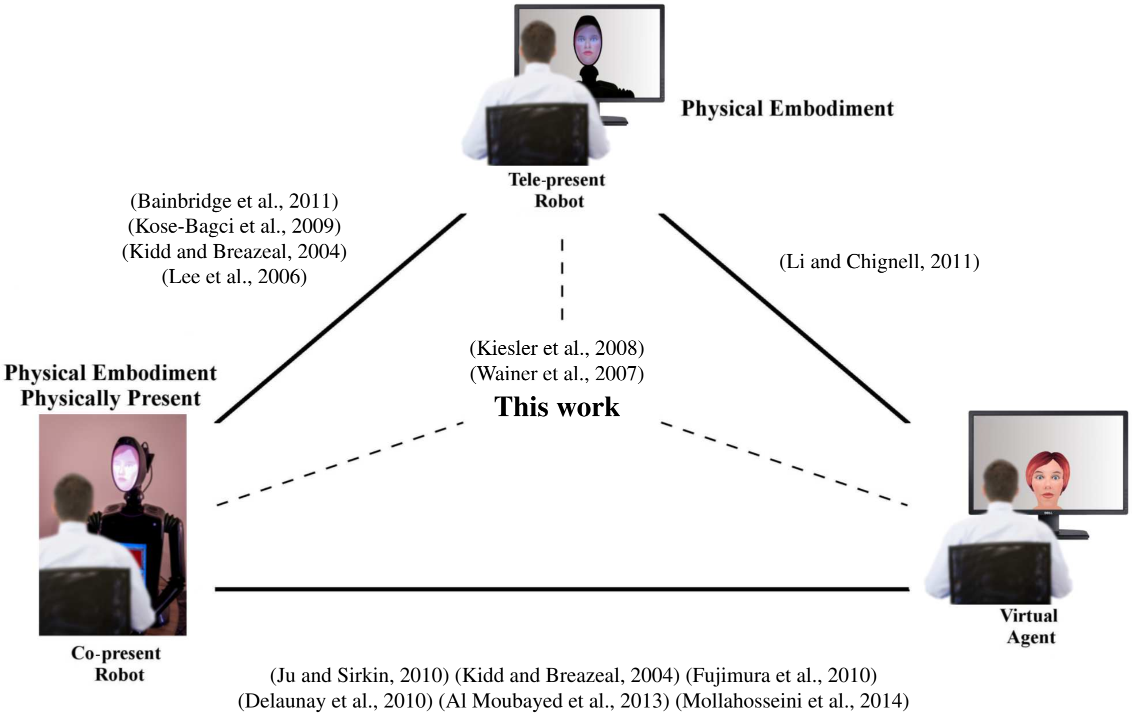

Physical Embodiment
Tele-present Robot
(Bainbridge et al., 2011)
(Kose-Bagci et al., 2009)
(Kidd and Breazeal, 2004)
(Lee et al., 2006)
(Li and Chignell, 2011)
(Kiesler et al., 2008)
(Wainer et al., 2007)
This work
Physical Embodiment
Physically Present
Co-present Robot
Virtual Agent
(Ju and Sirkin, 2010) (Kidd and Breazeal, 2004) (Fujimura et al., 2010)
(Delaunay et al., 2010) (Al Moubayed et al., 2013) (Mollahosseini et al., 2014)

Figure 3.2: Comparison of presence and embodiment dimensions across three categories of experimental stimuli in the literature (inspired from Li (2015)). The majority of studies do not distinguish the telepresence of a robot (physical embodiment) from the copresence of a robot (physical presence).

the mask is static and the jaw and lip movements are only optical). We consider these limitations in this study.

The remainder of this chapter is organized as follows. Section 3.3 reviews the definition of physical embodiment and presence and then defines research questions of this study. Sections 3.4, 3.5, and 3.6 study the role of embodiment and presence in perception of a robot's visual speech, facial expression, and eye gaze, respectively. In each of these sections, a brief review of prior work, the algorithm used to generate the facial cues, the experiments and settings, and the results, as well as a brief discussion of the results are presented. Finally, Section 3.7 discusses the results and findings and concludes this chapter.

3.3 Embodiment and Presence

Socially Intelligent Agents (SIAs) are systems that are able to connect and interface to humans via the ability to show aspects of human-style social intelligence (Dautenhahn, 1998). These agents can have a wide range of forms, some of which have physical bodies (e.g. a robot) or virtual observable bodies/faces (e.g. an intelligent avatar), and some of which interact with others using only voice or text without having any appearance (e.g. Siri). Since body gesture and expressions play a crucial role in social interactions and communication (e.g., body language, head gesture, facial expressions, speech, etc.), researchers try to build SIAs that closely mimic the appearance, behavior, and social skills of human beings (Dautenhahn, 2001). The field of "embodied conversational agents" is an excellent example of this approach (Cassell, 2000).

Mimicking the appearance of humans in SIAs or "*tighter coupling of the* [human] *body to the interface*" (Biocca, 1997) is viewed as central for providing the embodiment to the agents. This embodiment can be both virtual (e.g., embodied conversational virtual agents) and physical (e.g., robot). Pfeifer and Scheier (1999) defined the *physical embodiment* in intelligent robots as "a term used to refer to the fact that intelligence cannot merely exist in the form of an abstract algorithm but requires a physical instantiation, a body."

In-line with this definition, much work has examined the role of embodiment with regard to a variety of social interaction elements such as persuasion (Ju and Sirkin, 2010), likeability (Kidd and Breazeal, 2004; Kiesler et al., 2008), enjoyment (Wainer et al., 2007), trustworthiness (Kidd and Breazeal, 2004), helpfulness (Wainer et al., 2007), direct gaze recognition (Ju and Sirkin, 2010), and ease of interaction (Fujimura et al., 2010). The majority of these reports claimed that the physicality of the robot benefited user interaction. However, many of these studies did not distinguish physical embodiment from the copresence of the robot.

Copresence is a sociological concept describing the condition in which human individuals interact with each other (Goffman, 1963; Zhao, 2003). In our case, copresence refers to how the agent is presented to the user. Zhao (2003) defined copresence in two dimensions: 1) the mode of being with others (i.e., physical conditions that structure human interaction), and 2) the sense of being with others (i.e., subjective experience of being with others). The mode of copresence is related to the concept of "distance" in the taxonomy of copresence, which can be physical proximity (within range of the naked senses) or electronic proximity (outside the range of the naked senses but within the range of senses extended through electronic media) (Li, 2015). In real-world environments, physical and digital presence correspond to "copresence" and "telepresence," respectively (Zhao, 2003). The mode of copresence is also similar to the concept of "directness" in the literature (Li, 2015; Milgram et al., 1995). Physical and digital presence can be simply defined as a situation in which the embodied agent can be touched (or can touch the person). In other words, as Milgram et al. (1995) stated: *"[Physical or digital presence:] [the condition] whether primary world objects are viewed directly or by means of some electronic synthesis process."*

The mode of copresence (e.g., physical or digital) can affect a person's sense of copresence or "social presence" (Zhao, 2003). Some researchers evaluated the role of presence by comparing a robotic agent with its telepresence or a video of the robot. In a recent survey (Li, 2015), the effects of physical embodiment and physical presence were explored through a study of 33 experimental works to compare how people interact with 1) physically present robots, 2) telepresent robots, and 3) virtual agents. The study showed that physical presence plays a greater role in determining a person's response to an agent than physical embodiment. The methods used in these studies include post-treatment questionnaires or measuring subjects' behaviors during laboratory experiments. Among these 33 studies, however, few compared all three conditions in the same experiment/platform (See Fig. 3.2).

3.3.1 Research Questions

Based on the above and since face-to-face interaction is one of the essential elements of a social system, we have designed three research questions to be addressed in this dissertation:

- Q1: What is the effect of physical embodiment on perception of agents' facial cues (telepresent robot vs. virtual agent)?
- Q2: What is the effect of physical presence on perception of agents' facial cues (copresent robot vs. telepresent robot)?
- Q3: What is the joint effect of physical embodiment and presence on perception of agents' facial cues (copresent robot vs. virtual agent)?

In order to answer these research questions, we studied three major facial cues (i.e., visual speech, facial expressions and eye gaze) in this investigation. Each experiment included four conditions:

1. **Virtual Agent (VA):** An animated face was presented on a 2D screen.
2. **Copresent Robot (CR):** The robot was physically present in front of each subject.
3. **Telepresent Robot (TR):** A video or still image of the robotic head was presented to each subject. The videos/images were captured in a frontal angle of the physical agent, and the face in the video was scaled to match the size of the copresent robot.

4. **Human Ground-Truth (GT):** A human performed the task instead of the agent in front of each subject, or the subject was presented with a video recording of the human. If a video was presented, the size of the face in the video was scaled to match the size of the virtual agent's face. The purpose of performing the experiments with GT (human) is to evaluate what we expected to be optimal perception of social cues in our research setting.

In all four conditions, subjects were seated in front of the agent, with the same viewing angle and distance between the subjects and the agent. We used a rear-projected robotic head for this study since computer graphic generated avatars can show natural visual speech and facial expressions, and the same virtual agent animation behaviors can be compared with telepresent and physically present robotic agents. Similar to other robotic platforms, rear-projected robots have some limitations. For instance, Android robotic heads are limited by the number of actuators used in their face, or non-humanoid robots may not be able to show facial expressions. Similarly, since the mask is static in rear-project robotic heads, the jaw and lip movements are only optical and some facial movements (such as nose wrinkling during the expression of disgust) cannot be shown. Therefore, the findings of this investigation cannot be generalized to all other embodiments without considering the relevant differences between the embodied agents.

3.4 Visual Speech

Visual speech includes the visible oral cues (e.g., movement of the lips, tongue, and jaw) during speech production. These visual cues are not simply a by-product of speech production; they influence auditory perception of speech and vice versa. For example, McGurk and MacDonald (1976) showed that perception of mouth movements can affect the auditory perception of speech and Sweeny et al. (2012a) showed that hearing speech sounds influences the perception of simple visual shapes.

Considering the importance of speech and dialogue in social interaction, it is not surprising that many social robotic platforms have the capability of showing lip synchronization with auditory speech. Mechanical and Android robotic platforms such as Kismet (Breazeal, 2000), HRP-4C (Kajita et al., 2011), FR-i (Oh et al., 2010), Luo Head (Luo et al., 2011) and Alex (Lin et al., 2013a) have relatively basic visual speech due to limited actuators and mechanical components that are necessary to control the jaw movements. Computer graphic animations, on the other hand, have a greater capability for depicting natural visual speech, since mechanical actuators do not control the lips/jaw movements. Nevertheless, lack of physical embodiment and physical presence may constrain the perception of speech in graphic animations.

3.4.1 Related Work

Studies show that virtual embodied talking agents enhance the level of engagement, increase speech comprehension in noisy environments, make agents appear more life-like, and users tend to spend more time with these systems (Lester et al., 1999; Walker et al., 1994). Siciliano et al. (2003) compared SynFace Virtual Agent (VA) with audio (without visual speech) and video of a human, and concluded that visual-based speech intelligibility of this virtual agent is better than audio only, whereas it is significantly lower than audio-visual intelligibility of human visual speech. Ouni et al. (2003) performed a similar experiment on Baldi virtual agent. They eliminated syntactic and semantic cues by evaluating the perception of visual speech on a non-meaningful series of three Arabic words, and they concluded that speech is better perceived on VA with visual speech than with auditory

Table 3.4: Summary and overview of literature comparing audio-visual speech in different conditions

Work	Agent	Condition*				Description	Results**
		CR	TR	VA	GT		
Siciliano et al. (2003)	SynFace			✓	✓	• 12 normal hearing (NH) and 13 hearing-impaired (HI) listeners • Audio signal was degraded for NH group • Video of the original talker was used for GT	• Average intelligibility of VA increased by 22% compared to audio only • Intelligibility of VA was significantly lower than GT
Ouni et al. (2003)	Baldi			✓	✓	• Non-meaningful series of three Arabic words presented to 19 participants • Total of 300 words and 100 trials • Audio signal was degraded	• Average intelligibility of VA increased by 24% compared to audio only • Intelligibility of VA was 15% lower than GT
Al Moubayed et al. (2013)	Furhat	✓		✓	✓	• Audio-visual perception viewed at frontal and 45° angle. • A collection of short Swedish sentences • Reduced audio signal quality	• Audio-visual speech was better perceived on CR compared with VA. • No significant difference between frontal and 45° view angle
Mollahosseini et al. (2014)	Expressionbot	✓		✓		• Two short segments of speech • Examined two different lip synchronization approaches. • Participant rated how realistic the visual speech looked on a scale from 0 to 5	• Significant preference for the proposed visual speech approach over basic method • No significant preferences between CR and VA
This Work	Ryan	✓	✓	✓	✓	• Section 3.4.2	• Section 3.4.4

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.

** Only the relevant finding from the original papers are reported in this summary.

information only, but still nevertheless significantly lower than audio-visual intelligibility of human visual speech.

Only a few studies have compared the role of embodiment and presence of robotic agents in audio-visual speech perception. Al Moubayed et al. (2013) investigated the role of embodiment of a copresent robotic agent for improving the perception of visual speech. A facial animation on a 2D screen was compared with a rear-projection of the same animation using Furhat (Al Moubayed et al., 2012) and a video of humans from different viewing angles. A collection of short and everyday Swedish sentences with a length of three to six words in each sentence was created. The audio signal quality was reduced using band-pass filtering in specified frequencies and replaced with white noise. Six conditions were studied: audio only, virtual agent viewed at frontal and 45° angle, copresence of a robot viewed at frontal and 45° angle, and the original video recordings of the sentences viewed at the frontal angle. Fifteen sentences were examined in each condition. Auditory-visual perceptual sensitivity was measured as the number of correctly recognized words divided by the number of words in each sentence. This study, conducted on ten subjects with normal hearing, showed that audio-visual speech intelligibility was better perceived with the copresent robot (even though the jaw did not move in the mask), compared with the virtual agent on a flat screen. However, there was no significant difference in the audio-visual intelligibility of the face when it was looked at either from a front-view or a 45° angle on both the virtual agent and robot.

Mollahosseini et al. (2014) studied individuals' experiences and impressions of a proposed visual speech algorithm. In particular, they compared judgments of speech production quality of a virtual agent with rear-projection of the same animation using ExpressionBot. Two short segments of speech were presented with two different lip synchronization approaches (i.e., a proposed approach with kernel smoothing and lip closure in labial phonemes and a basic approach without any further smoothing and processing). The participants (23 typical adults) rated how realistic the visual speech looked on a scale from 0 to 5. Results showed a significant preference for the proposed lip synchronization approach over the basic approach. However, there was no difference in preference for visual speech from the virtual agent compared with the copresent robot.

Table 3.4 summarizes the results of studies on audio-visual speech intelligibility. As shown, none of these studies compared all three conditions of CR, TR, and VA to distinguish the role of embodiment from the presence of an intelligent agent in perception of visual speech. In this dissertation, we studied the perception of visual speech from three different types of emotional agents (i.e. VA, TR, CR) as well as from a human (as the optimal case) and based on auditory information alone (as the baseline) using the same experimental setup. Since the methodology and evaluation metrics of evaluating visual speech perception are not standard across the literature, we introduced a new test of visual speech perception along with standard criteria to evaluate the visual speech perception.

3.4.2 Methodology

We used the same visual speech algorithm presented in Chapter 2, Section 2.2, which is based on a multi-target morphing method (Ma and Cole, 2004). In particular, the recorded utterances are processed by the Bavieca speech recognizer (Bolanos, 2012), which receives the sequence of words and the speech waveform as input and provides a time-aligned phonetic transcription of the spoken utterance. The aligned phonemes are represented using the International Phonetic Alphabet (IPA), a standard that is used to provide a unique symbolic notational for the realization of phonemes in all of the world's languages (IPA-Handbook, 1999). Having IPA in our system will allow us to add other languages easily as long as the speech recognizer is trained for that language.

For a given language, visually similar phonemes are grouped into units called visemes. For example, the consonants /b/, /p/ and /m/ in the words "buy," "pie," and "my" form a single viseme class. English phonemes are categorized into 20 viseme classes. These classes represent the articulation targets that the lips and tongue move toward during speech production. A graphic artist designed 3D models of these viseme classes in Maya. Finally, natural visual speech was obtained by blending the proper models corresponding to each part of speech with different weights.

The avatar system converts phonetic symbols into the corresponding visemes, and synchronizes them with the audio signal. To achieve a smooth and realistic appearance, the algorithm models coarticulation by smoothing across adjacent visemes using a kernel technique, while ensuring lip closure for labial phonemes (e.g., /b/, /m/, /p/).

3.4.3 Visual Speech Experiment

Unlike with auditory speech (e.g., an evaluation of hearing ability), there is not a standard methodology to evaluate the perception of visual speech. Several researchers have thus developed their own approaches and evaluation criteria. The sets of sentences in the majority of these studies (See Table 3.4) are not comprehensive and do not consider syntactic and semantic cues. Measures of performance such as the number of correctly recognized words divided by the number of words in each sentence (Al Moubayed et al., 2013), or subjective evaluation of how realistic the visual speech appeared (Mollahosseini et al., 2014) are not standard, either. To address this issue, we developed an Audio-Visual Speech Perception In Noise (AV-SPIN) test to evaluate the perception of visual speech using a systematic and standardized approach. The AV-SPIN material, including videos, sentences, and IPA aligned auditory information, will be publicly available to the research community.[1]

The Speech Perception In Noise (SPIN) test was developed to address sensory and linguistic cognitive processes of everyday speech (Elliott, 1995; Kalikow et al., 1977). SPIN consists of 250 meaningful sentences categorized as High-Predictability (HP) sentences and 250 non-meaningful sentences categorized as Low-Predictability (LP) sentences. The listener's task is to recognize the last word in each sentence (referred to as the keyword). HP sentences contain syntactic and semantic cues helpful for predicting the keyword (e.g., *The sleepy child took a nap*), while LP sentences do not provide any cues predictive of the keyword (e.g., *Betty knew about the nap*). The sentences were divided into ten sets each containing 50 sentences (25 HP and 25 LP sentences), where odd-numbered sets were complementary of even-numbered sets (i.e., same keywords were in the opposite type of sentence).

Bilger et al. (1984) studied the SPIN test on 128 listeners (aged 19 to 69) with sensorineural hearing loss and proposed a revision (R-SPIN) such that different sets produce equivalent results. Particularly, 31 sentences and their complements were eliminated, 19 sentence pairs were arbitrarily removed, and the remaining sentences were redistributed to create 200 HP sentences and their complementary 200 LP sentences. These 400 sentences

[1] A copy of AV-SPIN is available in: http://mohammadmahoor.com/databases-codes/

were divided into eight sets each containing 50 sentences (25 HP and 25 LP sentences), where odd-numbered sets were complementary of even-numbered sets. Traditionally the R-SPIN is presented with ambient noise at a Signal-to-Noise Ratio (SNR) of 8 dB.

Since R-SPIN is strictly auditory, audio-visual intelligibility cannot be examined with the original R-SPIN materials. Therefore, we created an AV-SPIN corpus by capturing a native English speaker's face as she produced R-SPIN sentences. Similar to the R-SPIN test, the quality of the audio signal was degraded by babble noise. Since the subjects were not hearing-impaired, the audio signal was presented at a high signal-to-noise ratio of -9 dB (i.e., the power of the noise was significantly higher than the auditory speech signal).

Four conditions (VA, CR, TR, and GT), as well as audio only, were examined in the current experiment. In all conditions, subjects were seated in front of the agent at a distance of 60 cm. To maintain voice consistency between the conditions, the audio signals were extracted from the videos and force-aligned using Bavieca speech recognizer (Bolanos, 2012). Twenty sentences (10 HP and 10 LP sentences) were randomly assigned to each condition for each subject. A different set of sentences was used to train the subjects at the beginning of the experiment.

Each subject participated in all five conditions (audio-only, VA, CR, TR, and GT) in a random order. The LP and HP sentences in each condition were shuffled and were selected such that each condition did not share any sentences. The sentences were played only once, and at the end of each sentence, the subject had 30 seconds to write down the keyword (last word in each sentence). The subjects could adjust the sound volume at their convenience during the training period, but the same audio volume was used in all conditions of the remaining experiments. A set of headphones with the same audio volume was used in all the conditions. Since headphones were used, the direction of the voice did not play a role in the perception of speech. In addition, this allowed us to eliminate other roles, and only study the psychological effect of presence/embodiment of the robot. Only one trial was performed for each subject, since hearing a keyword in HP could have helped the subject to identify it in an LP sentence.

3.4.4 Visual Speech Results

Seventeen native English speakers with normal hearing listened to the audiovisual R-SPIN test corpus in five conditions. Figure 3.3 shows the mean accuracy for each condition. We performed a 2 (Predictability; HP, LP) × 5 (Condition; VA, CR, TR, GT, and Audio-Only conditions) ANOVA with both predictability and agent as the within-subject factors. The test showed a significant main effect of agent [$F(4, 64)$ = 30.48, p <.0001, h_p^2 = .656] and a significant main effect of predictability [$F(1, 16)$ = 134.55, p <.0001, h_p^2 = .894]. The interaction between agent condition and sentence predictability, however, was not significant [$F(4, 64)$ = 1.44, *n.s.*].

As Fig. 3.3 shows, VA (and all other conditions) produced significantly better audio-visual intelligibility than the audio-only condition [two-tailed t-test $p < .001$]. This confirms that visual information can affect speech perception, and shows the efficacy of the visual speech algorithm. The ground-truth (video of the human) had significantly higher audio-visual intelligibility than the other conditions [two-tailed t-test $p < .001$], which indicates that the proposed visual speech algorithm has room for improvement.

In order to measure the effect of agents' embodiment and presence in only VA, TR and CR conditions, we performed a 2 (Predictability; HP, LP) × 3 (Condition; VA, CR, TR) ANOVA with both predictability and agent condition as the within-subject factors. The analysis showed that the main effect of predictability was still significant [$F(1, 16)$ = 82.03, p <.0001, h_p^2 = .837], however the main effect of agent was not significant [$F(2, 32)$ = .381, *n.s.*] nor was the interaction between agent condition and predictability [$F(2, 32)$ = 1.44, *n.s.*]. In other words, embodiment and presence did not improve the perception

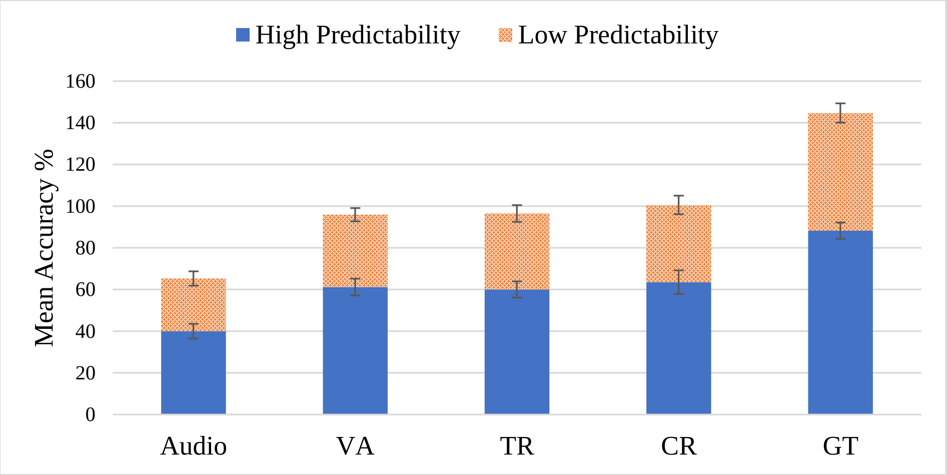


Figure 3.3: The average accuracy of audio-visual speech perception in different conditions. CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.

of visual speech regardless of syntactic and semantic cues in the sentences. A detailed analysis of these results and comparison of the findings with the literature are discussed in Section 3.7.

3.5 Facial Expressions

Facial expression is one of the most critical nonverbal channels used by human beings to convey emotion. Emotion is not only critical in creating more sensitive and effective intelligent agents but also impacts how people respond to the agent (Beer et al., 2011). Hence, facial expression is a vital component in natural social interaction and Human-Robot Interaction (HRI) systems, and has been employed in a variety of robots such as Kismet (Breazeal, 2003), the Philips iCat (Van-Breemen, 2004), Geminoid F (Becker-Asano and Ishiguro, 2011), and on-screen agents (Bruce et al., 2002; Cassell, 2000).

Mechanical and Android robotic platforms control face movement using actuators in their faces. However, due to cost and space constraints, the number of actuators in robotic faces are often limited. Moreover, because facial actions involved in facial expression can be very subtle and quick, mechanical actuators often fail to mimic them. Computer-graphic animations, on the other hand, have a greater capability for controlling facial movement, but their lack of physical embodiment and physical presence may constrain the perception of facial expression in virtual agents. Rear-projected robotic heads add physical embodiment to computer animation agents, but since the mask is static, some of the facial movements such as nose wrinkling in the expression of disgust cannot be portrayed on a robotic face. Therefore, it is important to investigate the role of embodiment and presence to find out whether physical embodiment and presence can improve the perception of an agent's facial expressions. A few studies have compared the role of embodiment and presence in the perception of robotic agents' facial expressions, and to the best of our knowledge perception of facial expression on rear-projected robotic heads has not yet been investigated.

3.5.1 Related Work

A few studies have compared the role of embodiment and presence in the perception of robotic agents' facial expressions. Bartneck et al. (2004) studied the role of presence in perception of intensity and recognition accuracy of facial expression using the robotic character iCat (Van-Breemen, 2004) and its telepresence condition (movie on a screen). Subjects were asked to categorize each emotion and rate its intensity. The study found a non-linear relationship between the geometrical intensity (robot's expression intensity) and the intensity of emotions perceived by the user. The results also indicated that emotions depicted by the robot were judged as having greater intensity, but there was no significant difference in the perceived intensity and recognition accuracy between the presence of the robotic character and its telepresence.

Kätsyri and Sams (2008) investigated the effect of dynamics on identifying basic emotions between a virtual agent (Talking Head) and a video of a human. Dynamic and static depictions of six basic emotions from a human face and a virtual agent were shown to 54 subjects. Subjects identified expressions on the human face much better than on the virtual agent. There was no significant difference in the identification of static and dynamic expressions of the human face. Identification of some expressions such as anger and disgust on the virtual agent failed to exceed chance level in the static condition, while dynamics improved it notably in lower intensities.

Table 3.5: Summary and overview of literature comparing perception of emotion in different conditions

Work	**Agent**	**Condition***				**Emotion†**		**Description**	**Results****
		CR	TR	VA	GT	No.	In		
Bartneck et al. (2004)	iCat	✓	✓			5	✓	• Ten geometrical intensities were displayed • Participants recognized the emotion and its intensity	• The relationship between the geometrical and perceived intensity was not linear • No significant difference between CR and TR in the intensity and recognition accuracy
Kätsyri and Sams (2008)	Talking Head			✓	✓	6		• Dynamic and static facial expressions were studied	• GT perceived better than VA • Dynamics did not improve GT • Dynamics improved recognition of subtle emotions, notably anger and disgust of VA.
Mollahosseini et al. (2014)	Expressionbot	✓		✓		6		• Participants selected six categories as well as "none"	• Superior recognition performance for anger in CR • Similar recognition rates for other emotions in both CR and VA
Lazzeri et al. (2015)	The Robot FACE	✓		✓	✓	6		• The robot, its 2D&3D models, and 2D&3D models of human were shown • Physiological signals of subjects were recorded	• CR was better perceived than 2D photos or 3D models (VA and GT) • No significant differences in the subjects' psychophysiological states
This Work	Ryan	✓	✓	✓	✓	6	✓	• Section 3.5.2	• Section 3.5.4

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.
†*No.* is the number of studied emotions and *In* stands for whether different Intensity levels are studied.
** Only the relevant finding from the original papers are reported in this summary.

Mollahosseini et al. (2014) studied the extent to which embodiment and physical presence improved the perception facial expression. The study evaluated how accurately subjects were able to interpret the facial expressions of a virtual 2D agent and its projection on a rear-projected robotic platform. Six basic emotions at their maximum intensity level were displayed in random order, and subjects were then asked to associate each with one of these six categories or to indicate that none were appropriate. They found similar recognition rates for happiness, sadness, surprise, disgust and fear in both a virtual agent and a copresent robot, and superior performance for anger when portrayed by the robot.

Lazzeri et al. (2015) studied the role of embodiment in conjunction with presence on a humanoid Android robot (Robot FACE). Fifteen subjects identified six basic emotions displayed on the robot, in 2D photos of the robot, 3D virtual animation models, as well as a set of 2D photos and 3D models of a human taken from Bosphorus Database (Savran et al., 2008). Preliminary results showed that facial expressions were better identified on the robot than its virtual animation, and the recognition rates of facial expressions performed by the robot were similar to those achieved with human stimuli.

Table 3.5 summarizes studies of facial expression perception with robots and their relevant findings. As shown, none of these studies compared all three conditions of CR, TR, and VA to distinguish the role of the embodiment from the presence of the robot. In this dissertation, we studied all three different conditions of emotional agents (i.e. VA, TR, CR) as well as human facial expressions (as the optimal case) in the same experimental setup. We also investigated emotion perception at different intensity levels to study the effect of intensity level on perception of different agents' facial expression.

3.5.2 Methodology

In order to design realistic and standard facial expressions in our animation system, we used the same algorithm presented in Chapter 2, Section 2.2.1, which is based on the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). The FACS model is a well-known approach for quantifying affective facial behaviors, and describes all possible

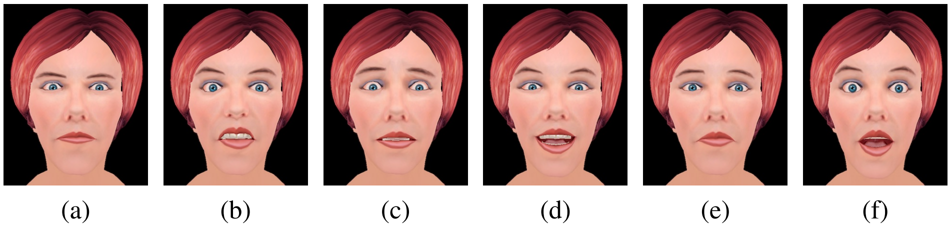


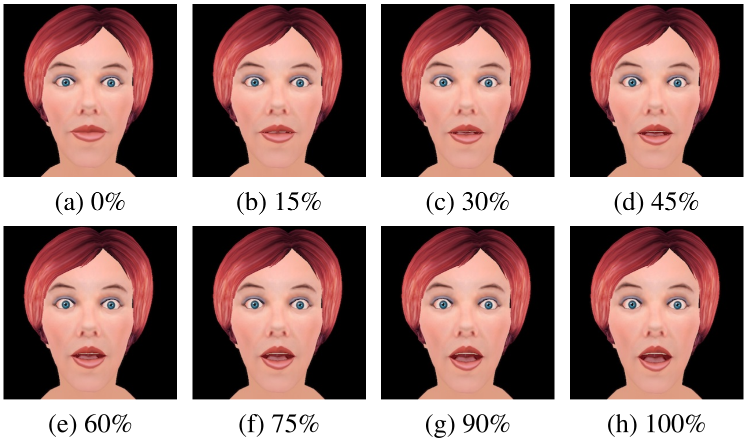
Figure 3.4: Six basic facial expressions at their maximum intensity: a) Anger, b) Disgust, c) Fear, d) Happiness, e) Sadness, and f) Surprise.

facial actions in terms of Action Units (AUs). The FACS explains facial movements and does not describe affective state directly. Friesen and Ekman (1983) proposed EMFACS to convert AUs to affect space. For example, EMFACS states that happiness involves raising of the cheek (AU 6) and pulling of the corner of the lip (AU 12), whereas sadness involves raising of the inner brow (AU 1), lowering of the outer brow (AU 4) and depression of the corner of the lip (AU 15). For the current experiment, a graphic artist designed 3D models of six basic expressions (i.e., anger, disgust, fear, happiness, sadness and surprise) in Maya based on EMFACS. Figure 3.4, demonstrates six basic facial expressions at their maximum intensity used in our animation system.

In order to show facial expressions at different intensities and blend them with visual speech, we used the same algorithm presented in (Mollahosseini et al., 2014). In particular, our animation used the following formula to generate the morph target based on the current viseme and emotion morph targets:

$$F_j = F_c + \lambda_j (F_j^{max} - F_0) \tag{3.1}$$

where F_c represents the current viseme, F_j^{max} is the desired expression model at the maximum intensity, F_0 is the Neutral model. The parameter $\lambda_j \in [0, 1]$ is the intensity of the j^{th} expression model F_j.



(a) 0% (b) 15% (c) 30% (d) 45%

(e) 60% (f) 75% (g) 90% (h) 100%

Figure 3.5: Different intensity level of surprised emotion on the virtual agent

3.5.3 Facial Expressions Experiment

Six basic facial expressions (anger, disgust, fear, happiness, sadness, surprise) were displayed in four conditions corresponding to the types of agents (VA, CR, TR, and GT) at seven intensity levels (15%, 30%, 45%, 60%, 75%, 90% and 100%). Each emotion was displayed with an animation/movie starting from neutral until the face's expression reached one of seven intensities. The animations took one second from neutral to the desired intensity and then remained static until the subject responded. Subjects were asked to categorize the emotion of the face as belonging to one of the six basic emotional categories (listed above) or to report "none" if they were unable to assign the facial expression to any of the six categories.

To evaluate the GT condition, subjects were presented with the video recordings of an actress portraying the facial expressions randomly selected from the extended CK+ dataset (Lucey et al., 2010). In order to pair the intensity of GT with the animation, two experts annotated the intensity of emotions between 0 to 100%, frame by frame. The intensity of each frame was considered as the average intensity of the two annotators. Each video in the GT condition took one second, started with a neutral expression, and ended at the desired emotional intensity level. Since the animation uses a weighted blend shape technique defined in (3.1), the intensity of emotion on the animation was easily defined by changing the parameter λ_j from zero to the desired intensity level over one second. Figure 3.5 shows different intensity levels of a sample emotion (surprise) on the virtual agent. Clearly, more subtle emotion intensities are more difficult to discriminate and could easily be confused with a different emotion.

Subjects were seated in front of the agent at a distance of 60 cm. Each combination of emotion and intensity was displayed twice in each block of trials, one with each intensity level, where the lowest intensity faces were shown first, then the second lowest, etc. In other words, subjects categorized 84 emotions (2 trials $\times$ 6 emotions $\times$ 7 intensities) where the first 12 videos/animations portrayed six emotions at intensity level 15% each played twice randomly, the second 12 videos/animations portrayed six emotions at intensity level 30%, and so on. The reason for sorting the trials by intensity level was that the subjects could have recognized the facial movement of an emotion at higher intensity levels and generalized the facial movements for recognition at lower intensity levels. In addition, each subject participated in only one agent condition, since VA, TR, and CR share the same animation and seeing an emotion at a higher intensity level of one condition could have helped the subject to recognize that same emotion at a lower intensity in another condition, on a different agent.

3.5.4 Facial Expression Results

We evaluated the perception of facial expressions of emotion performed by different agents with 48 subjects. Each subject participated only in one agent condition (i.e., 12 subjects rated the facial expressions displayed by one particular agent). In each condition,

Table 3.6: Confusion matrix of the emotion recognition rates (in percentage) of CR, TR and VA with presented facial expression (columns) against subjects' judgments (rows).

	Copresent Robot (CR)						**Telepresent Robot (TR)**						**Virtual Agent (VA)**					
	AN*	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU	AN	DI	FE	HA	SA	SU
Anger	**95.2**	2.4	0.6	0.0	0.6	0.0	91.1	5.4	0.0	0.0	1.8	0.0	81.5	4.8	0.0	0.0	0.6	0.6
Disgust	0.0	87.5	1.2	1.2	0.0	0.6	3.0	77.4	1.2	0.0	0.0	0.0	3.0	**90.5**	3.0	0.6	0.6	0.0
Fear	0.0	0.6	78.6	0.0	0.0	3.0	1.2	1.2	76.8	0.0	3.6	5.4	1.8	3.6	**81.5**	0.0	3.6	3.0
Happiness	0.0	0.0	0.0	**89.9**	0.0	1.2	0.0	0.6	0.6	83.9	0.0	0.6	0.0	0.0	0.0	89.3	0.0	0.6
Sadness	1.2	3.0	14.3	0.0	**98.2**	1.8	2.4	1.8	14.9	0.0	88.7	3.0	6.5	0.6	10.7	0.0	89.3	1.2
Surprise	0.0	1.2	2.4	7.7	0.0	91.1	1.2	4.2	3.6	10.7	0.0	81.5	0.6	0.6	4.2	10.1	0.0	**94.6**
None	3.6	5.4	3.0	1.2	1.2	2.4	1.2	9.5	3.0	5.4	6.0	9.5	6.5	0.0	0.6	0.0	6.0	0.0
Accuracy	**90.08**						83.23						87.80					

*AN, DI, FE, HA, SA, and SU stand for Anger, Disgust, Fear, Happiness, Sadness, and Surprise, respectively.

subjects saw facial expressions on an agent, and they were asked to select one of the six basic facial expressions (Anger, Disgust, Fear, Happiness, Sad and Surprise) or None.

A mixed 6 (emotions) × 7 (intensities) × 4 (agent conditions: CR, TR, VA and GT) ANOVA with emotion and intensity as the within-subject factors and embodiment as the between-subjects factor was conducted. The dependent variable was recognition accuracy. The recognition accuracy differed significantly between emotions [$F(5, 220)$ = 10.86, p <.0001, h_p^2 = .198] and between intensity levels [$F(6, 264)$ = 129.27, p <.001, , h_p^2 = .746]. Not surprisingly, faces with higher intensity received higher recognition accuracy. This analysis also revealed a significant interaction between the emotion and agent condition [$F(15, 220)$ = 1.95, p <.020, , h_p^2 = .117]. The interaction between agent and intensity was also significant [$F(18, 264)$ = 3.97, p <.001, h_p^2 = .213]. This suggests that the type of agent is particularly important when recognizing subtle expressions. The three-way interaction was also significant [$F(90, 1320)$ = 1.55, p <.001, , h_p^2 = .096]. This suggests that the dependency on intensity is only important for certain emotions (the intensity × agency interaction was significant for anger, fear, sad, and surprise, all p's <.05). Figure 3.6 shows the mean accuracy for each agent condition at different intensity levels, collapsed across the different expressions. As shown, the subjects discriminated emotion better on CR than on VA or TR.

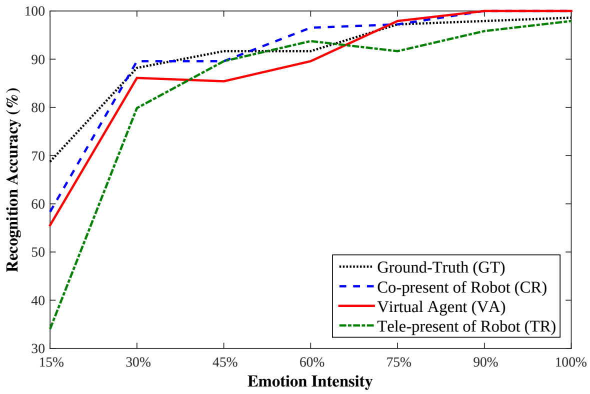


Figure 3.6: The average accuracy of emotion perception in different conditions.

There was also a significant main effect of agency on recognition accuracy [$F(3, 44)$ = 3.06, p = .038, h_p^2 = .173]. Post-hoc Least Significant Difference (LSD) analysis on different agent conditions indicated that expression recognition for TR was significantly worse than for human ground-truth and CR (p-values of 0.010 and 0.014, respectively). All other agent conditions were not significantly different from each other or ground-truth. In other words, both embodiment and presence were important factors in improving the perception of emotional expressions. Expression discrimination was better for the ground-truth (video of the human) condition than the other conditions, which indicates that the facial expression of the animation has room for improvement.

Table 3.6 shows the confusion matrices of the emotion recognition rates for the different agent conditions of CR, TR, and VA. The highest values are shown in bold. As shown, anger, happiness, and sadness were perceived better on CR, while disgust, fear, and surprise were recognized better on the virtual agent. To address whether this difference was significant between different emotions, separate post-hoc LSD analyses were conducted for each emotion. Table 3.7 shows the result of pairwise comparisons post-hoc LSD analyses and effect sizes of different agent conditions for different emotions.

Cohen's d is an effect size used to indicate the standardized difference between two groups defined as:

$$d = \frac{M_1 - M_2}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}} \tag{3.2}$$

where M_i is the mean and σ_i is the standard deviation of group i. Generally, the effect size is considered small if $d > 0.2$, medium if $d > 0.5$ and large if $d > 0.8$ (Cohen, 1977). As indicated in Tables 3.7:

- Anger was recognized better on both CR and TR compared to VA, with a medium effect size (effect of embodiment).

- Recognition of disgust, fear, and happiness was equivalent across all the agent conditions.

- Sadness was recognized better on CR compared to TR and VA, with a medium/large effect size (the effect of both embodiment and presence).

- Surprise was recognized worse on TR comparing with VA, with a medium effect size (effect of embodiment). However, Surprise was recognized better on CR compared with TR, with a medium effect size (the effect of presence).

A detailed analysis of these results and comparison of these findings with the literature are discussed in Section 3.7.

3.6 Eye Gaze

Eye gaze is one of the most basic and important features of the human face for nonverbal communication. Humans incorporate gaze both consciously and unconsciously into various human-human interaction schemes (Chen and Yeh, 2012). For example, neurons

Table 3.7: Pairwise comparison (LSD p-value) and Cohen's d effect size of users' perception of facial expressions on different agent conditions.

	TR vs CR		VA vs CR		VA vs TR	
	p	d	p	d	p	d
Anger	.405	.271	**.004**	**.652**	**.031**	**.419**
Disgust	.121	.347	.642	-.155	.050	-.493
Fear	.447	.139	.913	.002	.386	-.135
Happiness	.340	-.023	.784	-.230	.222	.218
Sadness	**.034**	**.789**	**.046**	**.727**	.891	-.008
Surprise	**.010**	**.421**	.426	-.237	**.001**	**-.658**

in the primate visual cortex can respond selectively to eye gaze, head orientation, or even the combination of both (Perrett et al., 1985). Eye gaze serves several different functions such as capturing attention, maintaining engagement (Cassell, 2000), conveying information about emotional and mental state (Ruhland et al., 2014), augmenting verbal communication (Emery, 2000), orchestrating turn-taking, and deictic reference (Kendon, 1967).

Considering the importance of eye gaze in social interaction, it is not surprising that social gaze behavior has been studied in many robotic platforms (Imai et al., 2002; Mutlu et al., 2009; Yoshikawa et al., 2006). Mechanical and Android robotic platforms control eye gaze by using actuators in the eyeballs. These actuators, however, may not be fast or accurate enough to replicate movement of the human eyes. The movement of the human eye is controlled by three pairs of muscles and it can reach an angular speed of about 400°/sec with 200ms time to initiate (Pateromichelakis et al., 2014). Computer graphics animations, on the other hand, have a greater capability for producing natural-looking eye gaze (Cassell, 2000; Ruhland et al., 2014). However, it is known that the perception of 3D objects that are displayed on 2D surfaces is influenced by the Mona Lisa effect (Todorović, 2006). Hence, the lack of physical embodiment and physical presence may constrain the perception of virtual agents' eye gaze.

3.6.1 Related Work

Table 3.8: Summary and overview of literature comparing perception of eye gaze in different conditions

Work	Agent	Condition*				EG†	Description	Results**
		CR	TR	VA	GT			
Anstis et al. (1969)	TV		✓		✓	✓	• A horizontal scale (ruler) was used • Video of a human used for TR • The agent's head was rotated with -30°,0° and 30° angles	• Errors were greatest when head rotation and eye rotation were incongruent.
Delaunay et al. (2010)	LightHead	✓	✓		✓	✓	• A grid with 100 cells was used • Video of a human used for TR • Instead of head rotation, subjects viewed the Agent with 0° and 45° angles	• CR performed better than TR • GT performed significantly better than other conditions, in both frontal and side view situations
Al Moubayed and Skantze (2012)	Furhat	✓			✓		• A grid with nine cells was used • Vergence, parallel eyes, static and dynamic eyelids	• Perception of gaze was significantly worse when the head was moving compared with eye movement alone. • No significant difference between gaze with and without vergence.
Moubayed et al. (2012)	Furhat	✓		✓			• Mona Lisa effect studied on five subjects sitting around a circle. • Only eye rotation studied	• Gaze was perceived more accurately on CR
Misawa et al. (2012)	LiveMask	✓		✓			• Photos of a person looking from -30° to 30° • Instead of rotating the head, subjects' view angle was changed	• CR was significantly better than VA • The Mona Lisa effect occurred in VR
Mollahosseini et al. (2014)	Expressionbot	✓		✓			• Mona Lisa effect studied on five subjects sitting around a circle	• Discrimination of eye gaze was better on CR
This work	Ryan	✓	✓	✓	✓	✓		

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.
†EG stands for Emergent Gaze which is defined as simultaneous movement of head and eye-gaze.
** Only the relevant finding from the original papers are reported in this summary.

Many studies in vision science have evaluated head-eye gaze, but only on telepresent faces (Allison et al., 2000; Baron-Cohen et al., 1995; Itier and Batty, 2009; Sweeny et al., 2012b). Although embodiment and presence have been studied individually, there is not a comprehensive study that distinguishes the role of embodiment and presence in gaze perception. Gaze perception of a physically present human agent and his video was studied on a TV set by Anstis et al. (1969). In this classic study, subjects were asked to report the point on a glass screen at which the agent (TV or a human) was looking. To simulate head rotation in the telepresent condition, the TV set was rotated. The agent's head was rotated to -30°, 0° and 30° angles. The study found that eye gaze was much better perceived on a physically present human agent than on its telepresent counterpart, and the perception of gaze was distorted with the rotation of the TV.

Delaunay et al. (2010) studied gaze perception on the LightHead robotic face, its telepresence, and the gaze of a human agent. A vertical glass screen with a 10x10 grid was placed between the agents and the subjects, and subjects were asked to report the gaze point when viewed from a frontal and 45° angle. Since asking a human to hold his/her head steady in a 45° position was not possible and chin/forehead rests did not allow horizontal rotations, to study the effect of head rotation, subjects were instead moved to a position with a 45° angle with respect to the agent. Under these conditions, subjects judged gaze from the video and the robot in both frontal and 45° view situations with equal sensitivity.

Al Moubayed and Skantze (2012) compared the perception of eye gaze on Furhat robotic face with a human agent in different conditions (i.e., presence of vergence, static/ dynamic eyelids, etc.). They took a different approach by asking the agents to look at nine points on a table between the agent and the subjects. In this case, there was no significant difference between gaze with vergence and without vergence. Furthermore, head movement appeared to be more effective for influencing judgments along the horizontal axis while eye movement dominated judgments along the vertical axis. Regardless of conditions, gaze from the human agent was perceived better than gaze from the robot.

Studies show that virtual agents suffer from the Mona Lisa effect (Misawa et al., 2012; Mollahosseini et al., 2014; Moubayed et al., 2012), in which the eyes in a picture appear to be looking at the viewer regardless of their location in front of the picture. For example, Moubayed et al. (2012) studied the Mona Lisa effect on a virtual agent and its 3D projection on Furhat robotic face. Five subjects were simultaneously seated around the agent, each of whom was asked to report their perception of the agents' eye gaze direction. The results showed a clear Mona Lisa effect in the virtual agent since many subjects perceived a mutual gaze at the same time.

Table 3.8 summarizes several studies on eye gaze perception and their most relevant findings. The majority of these studies report that physical presence plays a greater role in perception of an agent's eye gaze than physical embodiment. Presumably, having a 3D view of the nose direction, the eye position and their composition help viewers to perceive eye gaze direction more accurately. Additionally, few studies have explored emergent gaze. Emergent gaze occurs when the visual system integrates global information about the rotation of the head with local information about the rotation of the eyes, to compute a distinct metric of gaze present in neither feature alone (Cline, 1967; Kinya and Mitsuo, 1984; Kluttz et al., 2009; Langton et al., 2004; Otsuka et al., 2014; Sweeny and Whitney, 2017; Wollaston, 1824). This approach to measuring gaze perception has been surprisingly underutilized in robotics work.

The present study evaluates the perception of emergent gaze, while at the same time comparing the roles of embodiment and presence of the robot. One of the reasons that emergent gaze has not been studied extensively both with humans and robots is the difficulty inherent in controlling the movements of a human agent. Rotating a human's head and eyes to an exact position requires special apparatuses, and it complicates the experiment process. Hence, most studies of gaze either do not include a condition with a human agent, or they use a typical chin/forehead rest to fix the human's head in place, which precludes examination of emergent gaze.

3.6.2 Methodology

To evaluate the accuracy of agents' eye gaze in the current investigation, the agent looked at a particular point on a glass divider located between the agent and the subjects. A horizontal line with fifty-one equidistant points was drawn on the glass. The agent looked at a point on the glass screen and subjects were asked to report their perception of the agent's gaze direction.

In order to precisely set eye gaze toward a target point, we needed to rotate the agents' eyeballs such that the pupils were directed towards the target point. In this study, the target points were at agent's eye level, hence we only needed to change the yaw angle for the eyes. Assuming the face is frontal (rotated zero degrees), the yaw angle for right and left eyes (α_r and α_l, respectively) is calculated as:

$$\alpha_r = \frac{\pi}{2} - \arctan \frac{x + E_r}{D_r} \tag{3.3}$$

$$\alpha_l = \frac{\pi}{2} - \arctan \frac{x - E_l}{D_l} \tag{3.4}$$

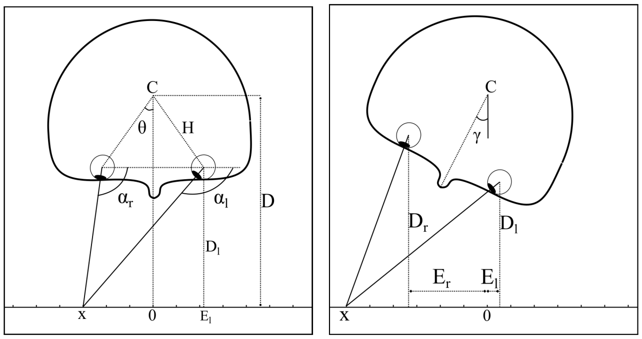
where $x \in [-75cm, 75cm]$ is the target point on the glass screen. E_r and E_l are the distance of right and left eye from the center of the glass screen in the x-Axis, and D_r and D_l are the distance of the right and left eyes from the glass screen in the y-Axis, calculated as:

$$E_r = E_l = H \times \sin(\theta) \tag{3.5}$$

$$D_r = D_l = D + H \times \cos(\theta) \tag{3.6}$$

where H is the distance of the head pivot point (C) to the center of the eyes, θ is the angle between the eyes and the head pivot point, D is the distance of the head pivot point to the glass screen. Figure 3.7a shows the schema and the variables used in these calculations.

When the head is straight and not rotated, $D_l = D_r$ and $E_r = E_l$. If the head is rotated by $\gamma°$ (Figure 3.7b), the values of E_r and D_r in Equations (3.3) and (3.4) are changed as



(a) Head facing forward (b) Head rotated by γ

Figure 3.7: Schema and the variables used in the calculating eye gaze angle (Drawing not to scale).

follows:

$$E_r = H \times \sin(\theta + \gamma) \tag{3.7}$$

$$E_l = H \times \sin(\theta - \gamma) \tag{3.8}$$

$$D_r = D - H \times \cos(\theta + \gamma) \tag{3.9}$$

$$D_l = D - H \times \cos(\theta - \gamma) \tag{3.10}$$

In the above equations, we assumed that the agent does not have any facial curvature in the eye area (Figure 3.8-left). If the face has an angle (ϵ) in the eye area (Figure 3.8-right), Equations (3.3) and (3.4) will change as follows:

$$\alpha_r = \frac{\pi}{2} - \arctan\frac{x + E_r}{D_r} - \epsilon \tag{3.11}$$

$$\alpha_l = \frac{\pi}{2} - \arctan\frac{x - E_l}{D_l} + \epsilon \tag{3.12}$$

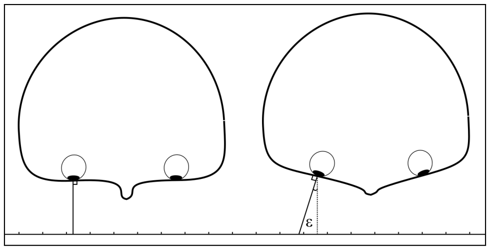


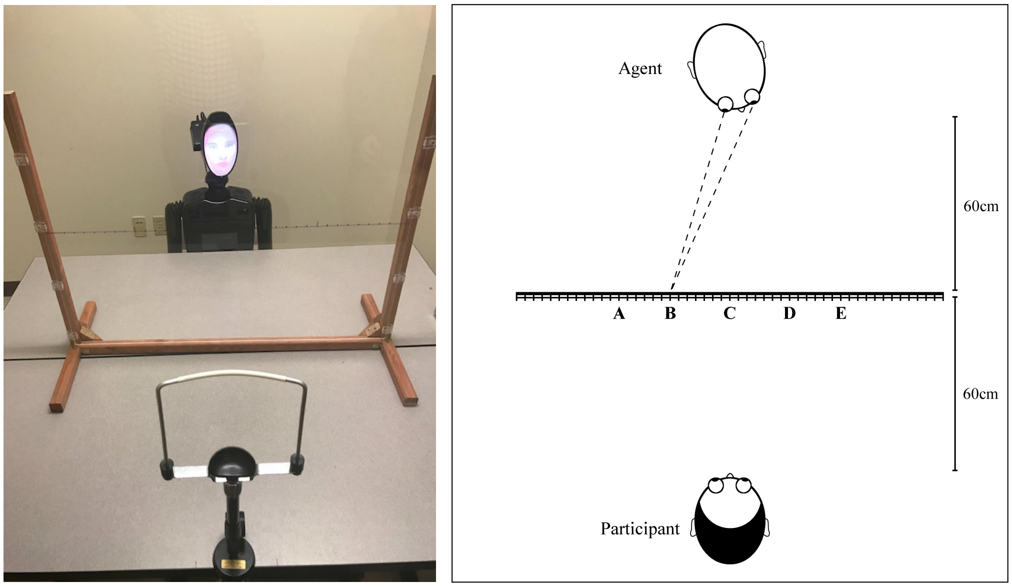
Figure 3.8: Mask with flat eye region (left) and with angled eye region (right)

3.6.3 Eye Gaze Experiment

To evaluate the role of embodiment and presence in perception of agents' eye gaze, four conditions (VA, CR, TR, and GT) were examined in this experiment. In each condition, the agent looked at a particular point on a glass divider located between the agent and the subjects. The subjects were then asked to report their perception of where the agent was looking.

The subjects were seated in front of the glass screen, and then asked to keep their head still on a chin-forehead rest and look straight at the agent at a distance of 120cm. To simulate the most accurate head rotation and avoid a Mona Lisa effect, which is common when viewing a face on a flat screen, in the VA condition we presented rotations of the animated head itself rather than rotations of the screen portraying the head. Figure 3.9 illustrates the eye gaze evaluation setup.

Fifty-one points, three centimeters apart from each other, were marked by letters and numbers on the glass. However, the agents looked at only five points located at -39, -21, 0, 21 and 39 centimeters (with zero as the middle point of the glass divider). Hereafter, these points are referred to as *A, B, C, D* and *E*, respectively (shown in Fig. 3.9). Subjects were not aware of the agent's restricted gaze targets, and they were instructed that the agent may look at any points on the glass. Figure. 3.10 shows photos of different conditions viewed from the subject's position.



(a) (b)

Figure 3.9: Perception of eye gaze room setup.

We examined the emergent perception of eye gaze (i.e., the integration of head rotation information with eye position). In particular, there were five possible head rotations (-30°, -16°, 0°, 16°, and 30°), and in each head position, the eyes were shifted toward the five points on the glass screen. An example of this condition is shown in Fig. 3.9, where the agent's head is rotated toward +16° and the eyes are directed at point *B*.

The method described in Section 3.6.2 was used to calculate the angle for the agent's eyes in CR and TR scenarios. The dimensions of the robot head for CR and the 3D model for VA were measured, and depending on the target point on the glass screen, the eyes of the robot/3D model were rotated toward the target point. The measurement used in CR was: $D = 73cm$, $H = 13.35cm$, $\theta = 13°$, and the measurement used in VA was: $D = 70cm$, $H = 10.45cm$, $\theta = 17°$. Since a mask with a flat eye region was used in CR and a flat screen was used in VA, the value of ϵ was set to $0°$.

A Canon EOS 80D DSLR camera was used to take pictures of the robot from the point of view of the subject. The captured pictures were calibrated to the size of the robot head.

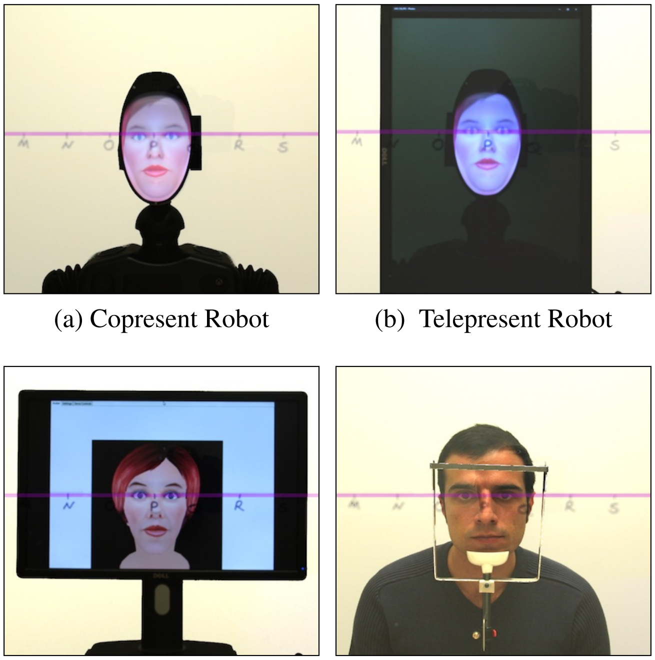
Using this method, from the point of view of the subject, the agent in both CR and TR had the same size and proportions, and in theory, the same direction of eye gaze (if we took a picture from the subject's point of view, it would look the same). The difference was that the TR condition featured a 2D representation of the CR condition.

To keep the human agent's head in an exact head rotation angle consistently during the GT experiments, we modified a chin/forehead rest to rotate and then stabilize in 1° increments. In the GT condition, a human was seated in the place of the agent and looked at the points on the glass, while keeping his head still on this chin forehead rest and his shoulders facing directly forward.

In all four conditions, first, the agent's head was rotated to one of the five angles (-30°, -16°, 0°, 16°, and 30°) randomly. Then at each of these head angles the eyes were rotated to gaze at one of the 10 points on the board (two trials for the five targets *A, B, C, D* and *E*) randomly. The subject was asked to close his/her eyes between each trial to eliminate any effect of seeing the agent adjust his head and eyes. In total, each subject reported 50 gaze directions (5 angles × 5 points × 2 trials) for each condition. Each condition was run in a block lasting five minutes and the subjects were asked to leave the room for two minutes until the room was setup for the next condition.

3.6.4 Eye Gaze Results

We examined the perception of eye gaze with 23 subjects, each of whom had normal or corrected to normal vision. Four different agent conditions (VA, TR, CR and GT) were presented in random order to the subjects, and subjects were asked to report their perception of the point at which the agent was looking. Accuracy was calculated by measuring the error in each subject's reports of eye gaze. Gaze perception error was defined as the



(a) Copresent Robot (b) Telepresent Robot

(c) Virtual Agent (d) Ground-Truth

Figure 3.10: Eye gaze different conditions

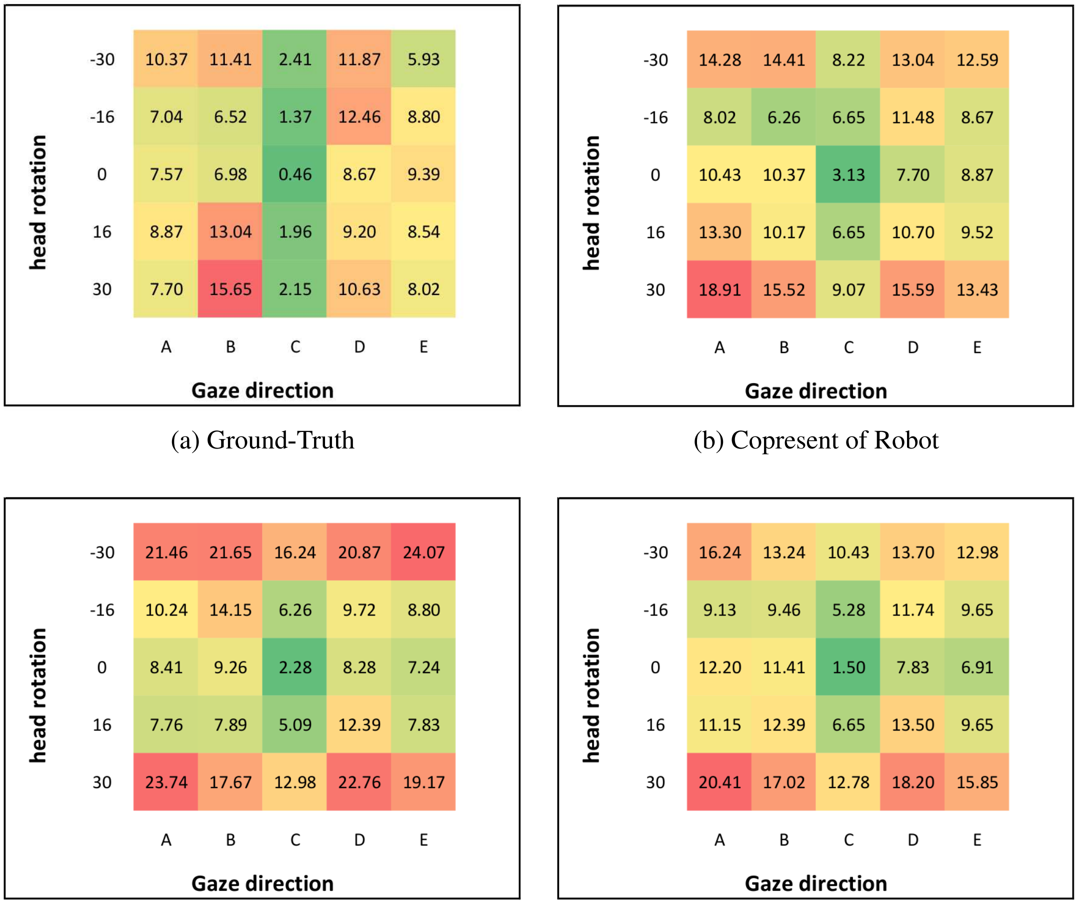
absolute distance between the point that the subjects reported and the actual target point at which the agent was looking.

We performed a 5 (head rotation) × 5 (eye gaze) × 4 (agent conditions: CR, TR, VA and GT) ANOVA with agent condition, head rotation and target point as within-subject factors. The dependent variable was gaze perception error. This analysis revealed a significant main effect of agent condition [$F(3,66)$ = 134.55, p <.0001, h_p^2 = .460]. We also found main

Table 3.9: Average and proportional error with respect to human ground-truth for different agent conditions.

	Average Error ± STD (cm)	Proportional Error to GT
GT*	7.88 ± 2.90	-
CR	10.50 ± 3.11	33.26%
TR	11.04 ± 3.16	46.47%
VA	13.04 ± 2.88	65.57%

* CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot,Virtual Agent, and Ground Truth (human) respectively.



(a) Ground-Truth

head rotation	A	B	C	D	E
-30	10.37	11.41	2.41	11.87	5.93
-16	7.04	6.52	1.37	12.46	8.80
0	7.57	6.98	0.46	8.67	9.39
16	8.87	13.04	1.96	9.20	8.54
30	7.70	15.65	2.15	10.63	8.02

(b) Copresent of Robot

head rotation	A	B	C	D	E
-30	14.28	14.41	8.22	13.04	12.59
-16	8.02	6.26	6.65	11.48	8.67
0	10.43	10.37	3.13	7.70	8.87
16	13.30	10.17	6.65	10.70	9.52
30	18.91	15.52	9.07	15.59	13.43

(c) Virtual Agent

head rotation	A	B	C	D	E
-30	21.46	21.65	16.24	20.87	24.07
-16	10.24	14.15	6.26	9.72	8.80
0	8.41	9.26	2.28	8.28	7.24
16	7.76	7.89	5.09	12.39	7.83
30	23.74	17.67	12.98	22.76	19.17

(d) Telepresent of Robot

head rotation	A	B	C	D	E
-30	16.24	13.24	10.43	13.70	12.98
-16	9.13	9.46	5.28	11.74	9.65
0	12.20	11.41	1.50	7.83	6.91
16	11.15	12.39	6.65	13.50	9.65
30	20.41	17.02	12.78	18.20	15.85

Figure 3.11: The average error (cm) in perception of different agents' eye gaze for different head rotation and gaze shift (Best viewed in color).

effects of head rotation [$F(4, 88)$ = 70.25, p <.0001, h_p^2 = .762] and eye gaze [$F(4, 88)$ = 31.39, p <.0001, h_p^2 = .588]. This analysis also revealed an interaction between agent condition and head rotation [$F(12, 264)$ = 11.17, p <.0001, h_p^2 = .337], but the interaction between the agent condition and eye gaze was not significant [$F(12, 264)$ = 95.16, *n.s*]. Figure 3.12 shows the estimated marginal means of gaze perception error for different agents, head rotation angle and target points. As shown, differences between the agent conditions depended on head rotation, but not eye gaze.

Table 3.9 shows the average and standard deviation of error for each condition and proportional error with respect to human ground-truth. The results indicate that eye gaze was better perceived on CR than TR and VA, with 13.21% and 32.23% lower proportional error, respectively. Figure 3.11 shows the average error (cm) in the perception of different agents' eye gaze for different head rotation and target points. As Fig. 3.11-(a) shows, when the eye gaze was directly toward the subject's face (point C), the perception of eye gaze had a relatively negligible amount of error. In other words, subjects were able to recognize mutual eye contact with high precision on the human agent. The same pattern emerged in the CR and TR conditions. Interestingly, subjects discriminated mutual eye gaze poorly in the VA condition, especially with incongruent head and eye rotations.

Notably, when the head was rotated to its extremes (-30° and 30°), perception of gazes directed toward points *B* and *D* had higher error than gazes directed toward points *A* and *E*. This suggests that subjects had difficulty recognizing gaze direction accurately when the rotation of the head was incongruent with that of the eyes. Hence, subjects may have guessed a point at the far end of the glass screen, which gave them more room for error at points *B* and *D*.

As shown in Fig. 3.11, eye gaze of the virtual agent was seen with a notable amount of error ($\sim$24cm) when combined with a strong head rotation. This could be because the animation lacked binocular depth cues by virtue of being present on a flat screen. This could have made the perception of head rotation more difficult, while the embodiment of the robot helped subjects to recognize the head angle better.

In order to more directly measure the effect of agents' embodiment and presence, we removed human GT from the analysis and performed a 5 (head rotation) × 5 (eye gaze) × 3 (agent conditions: CR, TR, VA) ANOVA with agent condition, head rotation and eye gaze as within-subject factors. This analysis revealed main effects of agent [$F(2, 44)$ = 8.740, p = .001, h_p^2 = .284], head rotation [$F(4, 88)$ = 64.95, p <.001, h_p^2 = .747] and eye gaze [$F(4, 88)$ = 16.39, p <.0001, h_p^2 = .427]. Similar to previous analysis, and as shown in Figure 3.12, there was a significant interaction between the agent condition and head rotation [$F(8, 176)$ = 8.75, p <.0001, h_p^2 = .285], but the interaction between the agent condition and eye gaze was not significant [$F(8, 176)$ = 23.98, *n.s*].

Since there was an interaction between the agent condition and head rotation, we performed pairwise two-tailed t-test comparisons between agent conditions at different head rotations. Table 3.10 shows pairwise p-value and Cohen's d effect-size between agent conditions. As shown, embodiment improved the perception of eye gaze at -30° and 30°, as indexed by significant differences between TR and VA conditions ($p < .001$ and $p = .023$ with large effect sizes $d = 1.22$ and $d = 0.69$ respectively). Physical presence did not improve the perception of eye gaze, as the differences between TR and CR conditions were not significant at any head angle. There were also significant differences between CR and VA at -30° and 30° (both $p < .001$ with large effect sizes $d = 1.49$ and $d = 0.89$ respectively). Because TR and VA were both significantly different at these head angles, we conclude that improvement in the perception of eye gaze compared with CR is mainly due to embodiment rather than presence of the robot. And in particular, embodiment of the robot highly affected the precision of the gaze perception combined with extreme head rotations in a frontal situated setting.

Table 3.10: Pairwise comparison (LSD p-value) and Cohen's d effect size of users' perception of eye gaze at different head rotations. Significant pairs are shown in bold.

	TR vs CR		VA vs CR		VA vs TR	
Head Angle	p	d	p	d	p	d
-30°	.660	0.13	**<.001**	**1.49**	**<.001**	**1.22**
-16°	.479	0.21	.190	0.39	.5484	0.17
0°	.890	-0.04	.278	-0.32	.269	-0.32
16°	.599	-0.15	.116	0.47	.158	0.42
30°	.217	0.36	**.004**	**0.89**	**.023**	**0.69**

3.7 Discussion and Conclusion

In this Chapter, we aimed to find the value propositions of a rear-projected robot compared with an on-screen animation. For this purpose, we performed two sets of HRI experiment. At first, individuals' experiences of interpreting the facial expressions and the proposed visual speech of ExpressionBot is compared with the facial animation on the computer screen. During these experiments, the users were in front of the robot, and it was not clear whether the users benefited from the physicality of the robot or they were under the impression of its physical presence. We then distinguished the role of the robot's embodiment from its physical presence in three major facial cues (i.e., visual speech, facial expressions and eye gaze). In particular, three different conditions (i.e., copresent of the robot, telepresent of the robot, and virtual agent) were studied to answer whether the embodiment of the robot has any interaction value proposition compared with an on-screen animation. In particular, aimed to investigate the effect of physical embodiment (Q1), physical presence (Q2), and the joint effect of physical embodiment and presence, on human perception of agents' facial cues (Q3). Three major facial cues (i.e., visual speech, facial expressions and eye gaze) were studied in this research. To study these effects, we leveraged three different agent conditions (i.e., copresent robot, telepresent robot, and virtual agent) as well as human ground truth to evaluate the optimal case in our settings. Below, we discuss the results of these three separate experiments.

Visual Speech: We found that speech from a virtual agent, and all other conditions, produced significantly better audio-visual intelligibility than speech from auditory information alone (Section 3.4.4). Nevertheless, the human ground-truth (video of a human) condition produced significantly higher audio-visual intelligibility than the other conditions. This confirms that visual information can affect speech perception, and hence the proposed visual speech algorithm has room for improvement. Audio-visual intelligibility was not significantly different across the agent conditions with no significant interaction between agent condition and predictability of the sentences.

Our results indicated that physical embodiment (Q1), physical presence (Q2), and the joint effect of physical embodiment and presence (Q3) did not differ in the extent to which they improved the perception of visual speech regardless of syntactic or semantic cues in the sentences. This could be because the mask was static and the jaw and lip movements were only optical in the rear-projected robotic platform. Other types of embodiment, such as Android robots, may express different behaviors. However, since controlling natural lip movement on Android robots necessitates several actuators and a very elastic skin, existing Android robotic faces may even perform worse than computer graphics animations.

This finding is consistent with our earlier study (Mollahosseini et al., 2014), but inconsistent with the study by Al Moubayed et al. (2013), though similar rear-projection technology with a static mask was used in both studies. It is unlikely that the results were influenced by different visual speech algorithms. It is more likely that the difference between Al Moubayed et al. (2013) and our finding is due to different audio-visual corpus and the intelligibility measurement criteria. The audio-visual corpus used in the present study was a standard set considering the syntactic and semantic cues in the sentences, while Al Moubayed et al. (2013) used a collection of short, everyday sentences with the number of correctly recognized words divided by the number of words in each sentence as the criterion of perception. Additionally, the sample size may also have affected the results, as the study performed by Al Moubayed et al. (2013) was evaluated with ten subjects, compared to this study that 17 subjects participated in.

Facial Expression: We found that expression recognition for the virtual agent and copresent robot were not significantly different from expression recognition from human ground-truth (Section 3.5.4). In other words, the facial expression generation algorithm can portray emotions similar to those from a human. Separate post-hoc LSD analyses for each emotion indicated that embodiment improved perception of the expression of anger (Q1), and embodiment and presence had a joint effect on improving perception of the expression of sadness (Q3). Physical embodiment impaired perception of the expression of surprise, however, physical presence could compensate for this negative effect (Q2).

We believe that the negative effect of physical embodiment on the perception of an agent's surprised expression could have occurred because the jaw does not move in the static mask, making subtly surprised faces difficult to perceive. This phenomenon (i.e., the effect of seeing a moving expression on a static mask) was presumably less noticeable when the robot was present in front of users (CR condition), as the difference between CR and VA was not significant for the expression of surprise. Since the only varying factor between TR and CR was the "presence" of the robot, we believe that presence could potentially compensate for the negative effect of seeing facial movements on a static mask.

These results are consistent with our previous study (Mollahosseini et al., 2014), indicating that subjects perceived the facial expression of anger (and sadness in the present study) with greater accuracy in the robotic face than that of the virtual agent. Our finding is also consistent with (Bartneck et al., 2004). We also found a significant difference between the robot and telepresence of the robot for perception of the facial expressions of sadness, similar to Bartneck *et al.*, who found a significant difference between CR and TR for recognizing sadness at intensities lower than 30%.

This finding is inconsistent with a study by Lazzeri et al. (2015) in which all emotions were better perceived on a robotic agent than on a virtual agent. Perhaps, the difference between (Lazzeri et al., 2015) and our finding is mainly due to the difference between the embodiments (i.e., Android vs rear-projected robotic heads). The masks in rear-projected robotic heads are static, thus jaw and the lip movements are only optical and some facial movements such as nose wrinkling in the expression of disgust cannot be shown, whereas Android robotic heads can be more flexible in controlling the skin if enough actuators are provided. In addition, Lazzeri et al. (2015) created a synthesized virtual agent from a set of pictures of a physical robot acquired from various angles and used Unity 3D software to animate the 3D models. Our virtual agent featured an accurate 3D model which was projected on the robotic face. Hence, the same animation and expression dynamics were used in both our robot and virtual agent conditions.

Eye Gaze: We found that there was a significant main effect of agent type between virtual agent, telepresent robot, copresent robot and human ground-truth (Section 3.6.4). There was a significant interaction between the agent condition and head rotation, but the interaction between the agent condition and the eye gaze was not significant. Eye gaze was better perceived on CR than TR and VA, with 13.21% and 32.23% lower proportional error. Pairwise comparisons between agent conditions at different head rotations showed that embodiment improved the perception of eye gaze at extreme head rotations (Q1). Physical presence did not improve the perception of eye gaze (Q2), as the difference between TR and CR conditions was not significant at any head rotation. There were also significant differences between the CR and VA conditions at extreme head rotations. Thus, because TR and VA were both significantly different at these head angles, improvement of eye gaze perception relative to CR was mainly due to embodiment rather than presence of the robot (Q3).

Table 3.11: Summary of role of embodiment and presence in perception of different facial cues.

	Physical Embodiment*	Physical Presence†	Join Effect**
Visual Speech	✗	✗	✗
Facial Expression	✓	✓	✓
Eye Gaze	✓	✗	✓

* Physical Embodiment indicates a significant difference between tele-present robot and virtual agent.

†Physical Presence indicates a significant difference between co-present robot and tele-present robot.

** Join Effect indicates a significant difference between co-present robot and virtual agent.

These findings are congruent with previous studies showing that the perception of a robot's eye gaze is more accurate than that of a virtual agent (Misawa et al., 2012; Mollahosseini et al., 2014; Moubayed et al., 2012). There was no difference in perception of gaze when seen on a robotic agent or its telepresence, which is consistent with a study performed by Delaunay et al. (2010). We also did not observe a significant difference between gaze perception on the telepresent robot and virtual agents—a comparison which has not been addressed in previous studies.

Table 3.11 summarizes the results of role of embodiment and presence in perception of different facial cues. The first column (physical embodiment) indicates a significant difference between tele-present robot and virtual agent. The second column (physical presence) indicates a significant difference between co-present robot and tele-present robot. The third column (join effect of physical embodiment and physical presence) indicates a significant difference between co-present robot and virtual agent.

3.7.1 Conclusion

The results of our initial HRI studies on a group of participants illustrated that the subjects perceived the facial expression anger with a much greater accuracy in the robotic face than the screen-based face and they also rated the generated visual speech smooth and realistic on both robotic and screen-based systems. In addition, we studied the perception of eye gaze's direction in two experiments, one in which the head was frontal and only the eye gaze was shifted, and the other with the head rotated but not necessarily correlated with the eye gaze direction. In both experiments, our results showed that participants perceived the robotic face mutual gaze more accurately.

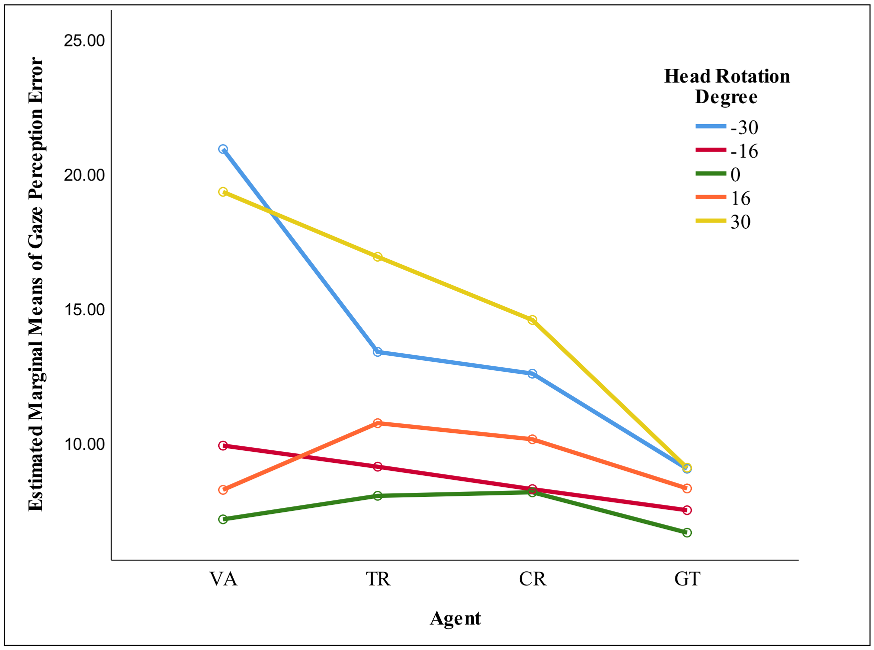
Then we examined the role of social robots' embodiment and presence in users' perception of facial cues using a quantitative approach. Understanding how people respond to physical and virtual agents is an important factor in designing successful social agents. The results of this study (summarized in Table 3.11) indicate that:

1. Neither embodiment nor presence plays a role at improving the perception of visual speech, regardless of syntactic or semantic cues in sentences.

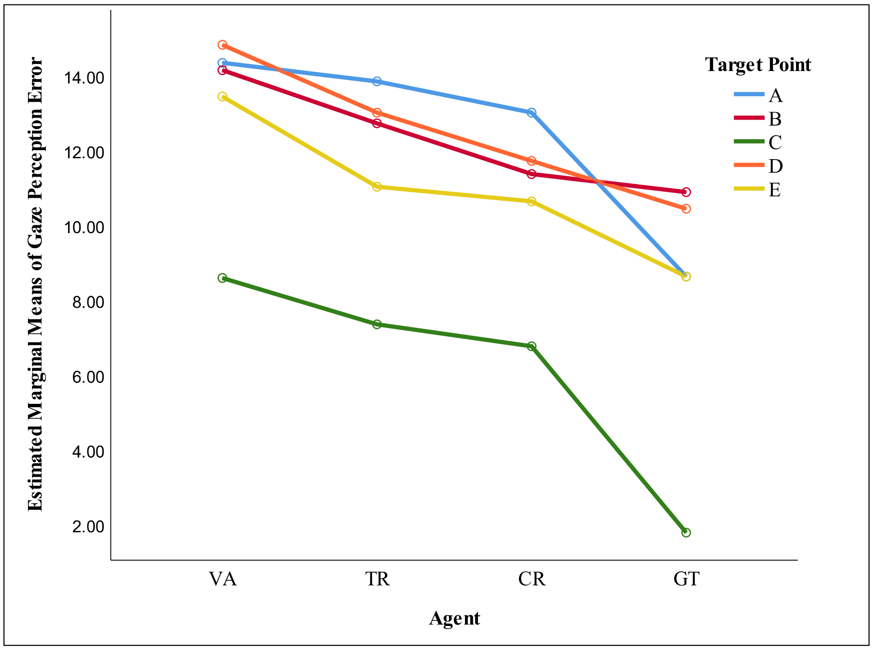
2. Both embodiment and physical presence improve the perception of certain facial expressions in emotive agents.

3. The combination of embodiment and presence (and mainly embodiment) highly affects the precision of eye gaze perception in a frontal situated setting.

Comparison of our findings with previous studies also indicates that the type of embodiment is important. We used a rear-projected robotic head in this study, which has some limitations (e.g., the mask is static, the jaw and lip movements are only optical). We believe that the limitations of embodiment can highly affect the perception of social cues. For instance, Android robotic heads are limited by the number of actuators used in their face and non-humanoid robots may not be able to show certain facial expressions. Therefore, the findings of any investigations on the role of embodiment and presence cannot necessarily be generalized to other types of robotic embodiments, without considering the characteristics of the embodied agents.



(a) Head rotation



(b) Target Point

Figure 3.12: Estimated marginal means of gaze perception error for different agents and (a) head rotation angles and (b) different gaze target points. The target points A, B, C, D corresponds to -39, -21, 0, 21 and 39cm from the center, respectively. CR, TR, VA, and GT stand for Copresent Robot, Telepresent Robot, Virtual Agent, and Ground Truth (human) respectively.

Chapter 4

Affect Perception

Current Human Machine Interaction (HMI) systems have yet to reach the full emotional and social capabilities necessary for rich and robust interaction with human beings. Facial expression, which plays a vital role in social interaction, is one of the most important nonverbal channels through which HMI systems can recognize humans' internal emotions. Ekman and Friesen (1971) identified six facial expressions (viz. anger, disgust, fear, happiness, sadness, and surprise) as basic emotional expressions that are universal among human beings.

Due to the importance of facial expression in designing HMI and Human Robot Interaction (HRI) systems (Mollahosseini et al., 2014), numerous computer vision and machine learning algorithms have been proposed for automated Facial Expression Recognition (FER). Also, there exist many annotated face databases with either human actors portraying basic expressions (Gross et al., 2010; Lyons et al., 1998; Mavadati et al., 2013; Pantic et al., 2005), or faces captured spontaneously in an uncontrolled setting (Dhall et al., 2013; Mavadati et al., 2013). Automated FER approaches attempt to classify faces in a given single image or sequence of images as one of the six basic emotions. Although, traditional machine learning approaches such as support vector machines, dictionary learning (Mohammadi et al., 2016) and to a lesser extent, Bayesian classifiers, have been successful when classifying posed facial expressions in a controlled environment, recent studies have shown that these solutions do not have the flexibility to classify images captured in a spontaneous uncontrolled manner (“in the wild”) or when applied databases for which they were not designed (Mayer et al., 2014). This poor generalizability of these methods is primarily due to the fact that many approaches are subject or database dependent and only capable of recognizing exaggerated or limited expressions similar to those in the training database. Many FER databases have tightly controlled illumination and pose conditions. In addition, obtaining accurate training data is particularly difficult, especially for emotions such as sadness or fear which are extremely difficult to accurately replicate and do not occur often real life.

Recently, due to an increase in the ready availability of computational power and increasingly large training databases to work with, the machine learning technique of neural networks has seen resurgence in popularity. Recent state of the art results have been obtained using neural networks in the fields of visual object recognition (Krizhevsky et al., 2012; Szegedy et al., 2014), human pose estimation (Toshev and Szegedy, 2014), face verification (Taigman et al., 2014), and many more. Even in the FER field results so far have been promising (Kahou et al., 2013). Unlike traditional machine learning approaches where features are defined by hand, we often see improvement in visual processing tasks when using neural networks because of the network’s ability to extract undefined features from the training database. It is often the case that neural networks that are trained on large amounts of data are able to extract features generalizing well to scenarios that the network has not been trained on. We explore this idea closely by training our proposed network architecture on a subset of the available training databases, and then performing cross-database experiments which allow us to accurately judge the network’s performance in novel scenarios.

The eventual goal of the proposed robotic platform is to interact with users in an uncontrolled setting (*"in the wild"*), where there is a high variation in scene lighting, camera view, image resolution, background, subjects head-pose and ethnicity. Since, existing facial expression recognition systems lack enough generality in the wild, we proposed a new Deep Neural Network (DNN) architecture. The proposed DNN is trained on seven well-known publicly available databases. However, the majority of these datasets are captured in a lab controlled setting. Therefore, we created a database of facial *Affect* from the InterNet (called **AffectNet**) by querying more than one million images from different search engines using 1250 emotion related tags in six different languages.

The rest of this Chapter is organized as follows: Section 4.1 reviews existing FER systems. Section 4.2 introdcues our proposed approach to extract facial landmark point, which is in a key step in facial image representation and analysis. Section 4.3 introduces the proposed DNN architecture, the process of learning from seven publicly available databases and the experimental results of subject-independent and cross-database settings. The process of creating and annotating *AffectNet* is explained in Section 4.4. Section 4.5 introduces two proposed DNN baseline to classify the facial expression images and predict the value of valence and arousal. At the end, Sec. 4.6 concludes the findings of this research on affect perception.

4.1 Existing FER Systems

Algorithms for automated FER usually involve three main steps, viz. registration, feature extraction, and classification. In the face registration step, faces are first located in the image using some set of landmark points during "face localization" or "face detection". These detected faces are then geometrically normalized to match some template image in a process called "face registration". In the feature extraction step, a numerical feature vector is generated from the resulting registered image. These features can be *geometric features* such as facial landmarks (Kobayashi and Hara, 1997), *appearance features* such as pixel intensities (Mohammadi et al., 2014), Gabor filters (Liu and Wechsler, 2002), Local Binary Patterns (LBP) (Shan et al., 2009), Local Phase Quantization (LPQ) (Zhen and Zilu, 2012), and Histogram of Oriented Gradients (HoG) (Mavadati et al., 2013), or *motion features* such as optical flow (Kenji, 1991), Motion History Images (MHI) (Valstar et al., 2004), and volume LBP (Zhao and Pietikainen, 2007). Since selecting the most appropriate feature is not trivial, Zhang et al. (2014, 2015) fuse multiple features using multiple kernel learning algorithms. However by using neural networks, we do not have to worry about the feature selection step - as neural networks have the capacity to learn features that statistically allow the network to make correct classifications of the input data. In the third step, of classification, the algorithm attempts to classify the given face as portraying one of the six basic emotions using machine learning techniques.

Cohn et al. (2007) distinguished two conceptual approaches to studying facial behavior: a “message-based" approach and a “sign-based" approach. Message-based approaches categorize facial behaviors as the the meaning of expressions, whereas sign-based approaches describe facial actions/configuration regardless of the action’s meaning. The most well-known and widely used sign-based approach is the Facial Action Coding System (FACS) (Ekman and Friesen, 1977). FACS describes human facial movements by their appearance on the face using standard facial substructures called Action Units (AUs). Each AU is based on one or a few facial muscles and AUs may occur individually or in combinations. Similarly, FER algorithms can be categorized into both message-based and sign-based approaches. In message-based approaches FER algorithms are trained on databases labeled with the six basic expressions (De la Torre and Cohn, 2011), and more recently, embarrassment and contempt (Lucey et al., 2010). Unlike message-based algorithms, sign-based algorithms are trained to detect AUs in a given image or sequence of images (De la

Torre and Cohn, 2011). These detected AUs are then converted to emotion-specified expressions using EMFACS (Friesen and Ekman, 1983) or similar systems (Taheri et al., 2014). In this research, we develop a message-based neural network solution,

FER systems are traditionally evaluated in either a subject independent manner or a cross-database manner. In subject independent evaluation, the classifier is trained on a subset of images in a database (called the training set) and evaluated on faces in the same database that are not elements of the training set often using K-fold cross validation or leave-one-subject-out approaches. The cross-database method of evaluating facial expression systems requires training the classifier on all of the images in a single database and evaluating the classifier on a different database which the classifier has never seen images from. As single databases have similar settings (illumination, pose, resolution etc.), subject independent tasks are easier to solve than cross database tasks. Subject independent evaluation is not, however, unimportant. If a researcher can guarantee that the data will align well in pose, illumination and other factors with the training set, subject independent evaluation can give a reasonably good representation of the classification accuracy in an online system. Another technique, subject dependent evaluation (person-specific), is also used in limited cases, e.g. FERA 2011 challenge (Valstar et al., 2011); often in these scenarios the recognition accuracy is more important than the generalization.

Recent approaches to visual object recognition tasks, and the FER problem have used increasingly "deep" neural networks (neural networks with large numbers of hidden layers). The term "deep neural network" refers to a relatively new set of techniques in neural network architecture design that were developed in order to improve the ability of neural networks to tackle big-data problems. With the large amount of available computing power continuing to grow, deep neural network architectures provide a learning architecture based in the development of "brain-like" structures which can learn multiple levels of representation and abstraction which allow algorithms for finding complex patterns in images, sound, and text.

It seems only logical to extend cutting-edge techniques in the field of “deep learning” to the FER problem. Deep networks have a remarkable ability to perform well in flexible learning tasks, such as the cross-database evaluation situation, where it is unlikely that hand-crafted features will easily generalize to a new scenario. By training neural networks, particularly deep neural networks, for feature recognition and extraction we can drastically reduce the amount of time that is necessary to implement a solution to the FER problem that, even when confronted with a novel data source, will be able to perform at high levels of accuracy. Similarly, we see deep neural networks performing well in the subject independent evaluation scenarios, as the algorithms can learn to recognize subtle features that even field experts can miss. These correlations provide the motivation for this reserach, as the strengths of deep learning seem to align perfectly with the techniques required for solving difficult “in the wild” FER problems.

A subset of deep neural network architectures called “convolutional neural networks” (CNNs) have become the traditional approach for researchers studying vision and deep learning. In the 2014 ImageNet challenge for object recognition, the top three finishers all used a CNN approach, with the GoogLeNet architecture achieving a remarkable 6.66% error rate in classification (Russakovsky et al., 2014; Szegedy et al., 2014). The GoogLeNet architecture uses a novel multi-scale approach by using multiple classifier structures, combined with multiple sources for back propagation. This architecture defeats a number of problems that occur when back-propagation decays before reaching beginning layers in the architecture. Additional layers that reduce dimension allow GoogLeNet to increase in both width and depth without significant penalties, and take an elegant step towards complicated network-in-network architectures described originally by Lin et al. (2013b). In other word, the architecture is composed of multiple “Inception" layers, each of which acts like a micro-network in the larger network, allowing the architecture to make more complex decisions.

More traditional CNN architectures have also achieved remarkable results. AlexNet (Krizhevsky et al., 2012) is an architecture that is based on the traditional CNN layered architecture - stacks of convolutions layers followed by max-pooling layers and rectified linear units (ReLUs), with a number of fully connected layers at the top of the layer stack. Their top=5 error rate of 15.3% on the ILSVRC-2012 competition revolutionized the way that we think about the effectiveness of CNNs. This network was also one of the first networks to introduce the "dropout" method for solving the over fitting problem (Suggested by Srivastava et al. (2014)) which proved key in developing large neural networks. One of the large challenges to overcome in the use of traditional CNN architectures is their depth and computational complexity. The full AlexNet network performs on the order of 100M operations for a single iteration, while SVM and shallow neural networks perform far fewer operations in order to create a suitable model. This makes traditional CNNs very hard to apply in time restrictive scenarios.

Liu et al. (2013) proposed a new deep neural network architecture, called an "AU-Aware" architecture was proposed in order to investigate the FER problem. In an AU-Aware architecture, the bottom of the layer stack consists of convolution layers and max-pooling layers which are used to generate a complete representation of the face. Next in the layer stack, an "AU-aware receptive field layer" generates a complete representation over all possible spatial regions by convolving the dense-sampling facial patches with special filters in a greedy manner. Then, a multilayer Restricted Boltzmann Machine (RBM) is exploited to learn hierarchical features. Finally, the outputs of the network are concatenated as features which are used to train a linear SVM classifier for recognizing the six basic expressions. Results in (Liu et al., 2013) show that the features generated by this "AU-Aware" network are competitive with or superior to handcrafted features such as LBP, SIFT, HoG, and Gabor on the CK+, MMI and databases using a similar SVM. However, AU-aware layers do not necessarily detect FACS defined action units in faces.

Kahou et al. (2013) combined multiple deep neural network architectures to solve the FER problem in video analysis. These network architectures included: (1) an architecture similar to the AlexNet CNN run on individual frames of the video, (2) a deep belief network trained on audio information, (3) an autoencoder to model the spatiotemporal properties of human activity, and (4) a shallow network focused on the mouth. The CNN is trained on the private Toronto Face Database (Susskind et al., 2010) and fine tuned on the AFEW database (Dhall et al., 2013), yielded an accuracy of 35.58% when evaluated in a subject independent manner on AFEW. When combined with a single predictor, the five architectures produced an accuracy of 41.03% on the test set, the highest accuracy in the EmotiW 2013 (Dhall et al., 2013) challenge, where challenge winner 2014 (Liu et al., 2014b) achieved 50.40% on test set using multiple kernel methods on Riemannian manifold.

A 3D CNN with deformable action parts constraints is introduced in (Liu et al., 2014a) which can detect specific facial action parts under the structured spatial constraints, and obtain the discriminative part-based representation simultaneously. The results on two posed expression datasets, CK+, MMI, and a spontaneous dataset FERA achieve state-of-the-art video-based expression recognition accuracy.

4.2 Facial Landmark Point Extraction

Facial landmark point extraction is a key step in facial image representation and analysis. The Active Appearance Model (AAM) proposed by Cootes et al. (2001) is a powerful object description method that is commonly used for facial landmark points extraction (Cootes et al., 2001; Matthews and Baker, 2004), facial action unit extraction (Mahoor et al., 2009), medical image segmentation and analysis (Cootes et al., 2004). The idea behind AAM is to represent a visual object (e.g. facial image) using a linear model of shape and texture (appearance) eigenvectors obtained from a set of manually labeled training images. Then, the model is used to represent an instance of the object in a novel image. This process is often called **AAM fitting**.

AAM fitting is a non-linear optimization problem. Different optimization approaches have been proposed to find the best model parameters that result in minimum error between the synthesized appearance models obtained from the AAM and the input image. In general, due to variation of camera view angle, resolution and focal distance, facial images have different scaling, rotation, and translations. In order to remove global shape variations, all shapes are normalized and the modeling is only concerned with local shape deformation. Therefore, it is necessary to combine a global shape transformation with the normalized AAM. The global shape transformation is often a 2D similarity transformation. Finding optimal parameters of the global transformation improves the accuracy of fitting in representing novel facial images with different shape and pose variations.

Traditionally, the stochastic gradient descent algorithm or iteratively incremental additive techniques are used to update the AAM parameters to fit onto novel images (Cootes et al., 2001). The fitting problem can also be viewed as finding a model instance similar to the given facial image and therefore it can be considered as an image alignment problem. Baker and Matthews (2004) have categorized these approaches into four classes: Forwards Additive, Forwards Compositional, Inverse Additive, and Inverse Compositional. Matthews and Baker (2004) proposed Projecting Out (PO) technique which is admittedly one of the fastest algorithms for AAM fitting. Gross et al. (2005) also proposed the Simultaneously Inverse Compositional (SIC) method that can handle images of subjects not included in the training better at the price of losing speed.

In the literature, there are some works (Keller and Averbuch, 2004; Mégret et al., 2010) on image alignment for applications, such as motion estimation (Keller and Averbuch, 2004), that take advantage of the gradients of both the template and target images. These approaches are called bidirectional image alignment. Bidirectional approaches work better than unidirectional image alignment approaches (Mégret et al., 2010). In this work, we reformulate AAM fitting using a bidirectional image alignment scheme.

In our approach, we minimize the error between a warped image and the appearance template by iteratively solving a non-linear least square problem. The warping is a piecewise affine of a normalized AAM that is followed by a global transformation. In each iteration, shape parameters are optimized based on the trained appearance template using the Inverse Compositional Algorithm (ICA) (Baker and Matthews, 2004), and global transformation is found based on the gradient of the input image using incremental update. We call this approach bidirectional warping. Moreover, we utilize affine transformation instead of 2D similarity to increase the generality of the global shape transformation, and apply a fitting constraint to prevent the algorithm from resulting in non-face shapes. We show that the proposed bidirectional approach can be applied to PO and SIC fitting methods. We study the performance of the proposed bidirectional PO and SIC methods in extracting facial landmark points, and examine and compare the effect of proposed affine transformation, and the fitting constraint on both bidirectional and the original PO and SIC fitting methods.

4.2.1 Active Appearance Model (AAM)

AAM consists of a shape component and an appearance component obtained from a set of annotated landmark points in training images. Let's assume we are given a training facial image set with annotated shapes defined as: $\mathbf{s} = (x_1, y_1, x_2, y_2, ..., x_v, y_v)^T$. The training images are first normalized and aligned using iterative Procrustes analysis (Cootes et al., 2004). This step removes variations due to a chosen global shape normalization transformation so that the resulting model can efficiently consider local and non-rigid shape deformation. We then can combine the resulting AAM with a global transformation. Afterwards, Principal Component Analysis (PCA) is applied to the set of normalized training shapes and a shape model is defined as:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i, \tag{4.1}$$

where the base shape $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are n eigenvectors corresponding to the n largest eigenvalues. Then, all the training images are normalized by warping them into the base shape $\mathbf{s}_0$, using piecewise affine warp, and the appearance model is defined as:

$$\mathrm{A}(\mathbf{x}) = \mathrm{A}_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \mathrm{A}_i(\mathbf{x}) \qquad \forall \mathrm{x} \in \mathbf{s}_0, \tag{4.2}$$

where A_0 is the mean appearance and the vectors A_i are the m eigenvectors corresponding to the m largest eigenvalues.

The goal of fitting is to find a model instance that can efficiently describe the object (e.g. face) in a given image. Thus, it can be considered as an image alignment problem. In other words, we want to find the model instance $M(\mathbf{W}(\mathbf{x};\mathbf{p})) = A(\mathbf{x})$ as similar as the image $I(\mathbf{x})$.

In general, facial images have different scaling, rotation, and translations. Therefore, it is necessary to combine a global shape transformation with the normalized AAM. If we consider the global shape transformation as $\mathbf{N}\left(\mathbf{x};\mathbf{q}\right)$, we want to minimize the error between the template and $I\left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x};\mathbf{p}\right);\mathbf{q}\right)\right)$. Considering global shape transformation, the objective of the fitting process is to find $\mathbf{p}$ and $\mathbf{q}$ in order to minimize the error image as:

$$E(\mathrm{x}) = \sum_{\mathbf{X} \in \mathbf{s}_0} \left[\mathrm{A}_0(\mathbf{x}) - I\left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x};\mathbf{p}\right);\mathbf{q}\right)\right)\right]^2, \tag{4.3}$$

which is a non-linear least square problem. We can have different definitions for the global transformation $\mathbf{N}\left(\mathbf{x};\mathbf{q}\right)$. Matthews and Baker (2004) defined a set of 2D similarity transformations as a subset of piecewise affine warps. Assuming the base mesh $\mathbf{s}_0 = \left(x_1^0, y_1^0, ..., x_v^0, y_v^0\right)^{\mathrm{T}}$, we choose $\mathbf{s}_1^* = \mathbf{s}_0$, $\mathbf{s}_2^* = \left(-y_1^0, x_1^0, ..., -y_v^0, x_v^0\right)^{\mathrm{T}}$, $\mathbf{s}_3^* = \left(1, 0, \ldots, 1, 0\right)^{\mathrm{T}}$ and $\mathbf{s}_4^* = \left(0, 1, \ldots, 0, 1\right)^{\mathrm{T}}$, then global transformation is $\mathbf{N}\left(\mathbf{x};\mathbf{q}\right) = \mathbf{s}_0 + \sum_{i=1}^{4} q_i \mathbf{s}_i^*$. This representation of $\mathbf{N}\left(\mathbf{x};\mathbf{q}\right)$ is similar to $\mathbf{W}\left(\mathbf{x};\mathbf{p}\right)$ and therefore similar analysis on the shape parameters $\mathbf{p}$ can be applied to $\mathbf{q}$. If we assume that the two sets of shape vectors $\mathbf{s}_i$ and $\mathbf{s}_i^*$ are orthogonal to each other, we can add the four 2D similarity vectors $\mathbf{s}_i^*$ to the beginning of AAM shape vectors $\mathbf{s}_i$ (Matthews and Baker, 2004) and model any given shape as: $\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n+4} p_i \mathbf{s}_i$. In practice, $\mathbf{s}_i$ and $\mathbf{s}_i^*$ are not quite orthogonal to each other. This can either be ignored when the size of $\mathbf{s}_i$ is small or the complete set of $\mathbf{s}_i$ and $\mathbf{s}_i^*$ can be orthonormalized preferably.

Baker and Matthews (2004) related AAM to the Lucas-Kanade algorithm. They proposed the Inverse Compositional Algorithm (ICA), in which they find shape variation on the template and compose the inverse of that with the current shape. Therefore, many computationally expensive tasks are precomputed.

Matthews and Baker (2004) considered appearance variation in the fitting by finding shape parameters in a linear subspace where the appearance variation is ignored and then "projected out" to the full space with respect to the appearance eigenvectors. The proposed method is more generic compared with the ICA, but the fitting is not accurate when applied to subjects that are not similar to subjects in the training set. The "projecting out" approach is called PO in the rest of this work.

Gross et al. (2005) introduced Simultaneously Inverse Compositional (SIC) method, which is more generic. In this method the fitting procedure minimizes the error between $[\mathrm{A}_0(\mathrm{x}) + \sum_{i=1}^{m} (\lambda_i + \Delta\lambda_i) A_i]$ and $I(\mathbf{N}(\mathbf{W}(\mathbf{x;p});\mathbf{q}))$, where A_i are m appearance eigenvectors correspond to the m largest appearance eigenvalues, and $(\lambda_i + \Delta\lambda_i)$ are parameters of appearance that are found simultaneously with respect to the $\Delta\mathbf{p}$. As the appearance parameters are optimized in each iteration, both steepest descent and the Hessian matrix (H) should be calculated in each iteration, and therefore the method is slower. Gross et al. (2005) compared PO with the SIC, and reported that SIC is more accurate in modeling unseen subjects.

4.2.2 Bidirectional Warping for AAM Fitting

In this work, we optimize the global transformation's parameters (**q**) based on I, using an incremental update and the shape's parameters (**p**) based on A_0, using inverse compositional approach. If we assume **p** and **q** are known, reversing the role of **W** in $I(\mathbf{N}(\mathbf{W}(\mathbf{x};\mathbf{p});\mathbf{q}))$ and computing the incremental global warp **N** with respect to **W** in $I(\mathbf{N}(\mathbf{W}(\mathbf{x};\mathbf{p});\mathbf{q}))$, we can solve the Equation (4.3) iteratively as:

$$\sum_{\mathrm{x}\in\mathbf{s}_0} [\mathrm{A}_0(\mathbf{N}(\mathbf{W}(\mathbf{x};0+\Delta\mathbf{p});0)) - I(\mathbf{N}(\mathbf{W}(\mathbf{x};\mathbf{p});\mathbf{q}+\Delta\mathbf{q}))]^2. \tag{4.4}$$

Then to update the warping parameters, we use $\mathbf{W}(\mathbf{x};\mathbf{p}) \leftarrow \mathbf{W}(\mathbf{x};\mathbf{p}) \circ \mathbf{W}(\mathbf{x};\Delta\mathbf{p})^{-1}$ and $\mathbf{q} = \mathbf{q} + \Delta\mathbf{q}$. Assuming $\mathbf{W}(\mathbf{x};0)$ and $\mathbf{N}(\mathbf{x};0)$ are identity warps, first order Taylor series expansion of the Equation (4.4) on $\Delta\mathbf{p}$ and $\Delta\mathbf{q}$ gives:

$$\sum_{\mathrm{x}\in\mathbf{s}_0} \left[\mathrm{A}_0 + \nabla\mathrm{A}_0\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\Delta\mathbf{p} - I(\mathbf{N}(\mathbf{W}(\mathbf{x;p});\mathbf{q})) - \nabla I\frac{\partial\mathbf{N}}{\partial\mathbf{q}}\Delta\mathbf{q}\right]^2, \tag{4.5}$$

where ∇ is the image gradient, $\frac{\partial\mathbf{W}}{\partial\mathbf{p}}$ and $\frac{\partial\mathbf{N}}{\partial\mathbf{q}}$ are the Jacobian of the warp evaluated at $\mathbf{p} = 0$ and current **q** respectively. By taking the derivative of the Equation (4.5), neglecting second order $\Delta\mathbf{p}\Delta\mathbf{q}$ terms and optimizing for $\Delta\mathbf{p}$ and $\Delta\mathbf{q}$, we obtain:

$$\Delta\mathbf{p} = \mathrm{H}_1^{-1}\sum_{\mathrm{x}}\left[\nabla\mathrm{A}_0\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^T\left[I\left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x};\mathbf{p}\right);\mathbf{q}\right)\right)-\mathrm{A}_0\right], \tag{4.6a}$$

$$\Delta\mathbf{q} = \mathrm{H}_2^{-1}\sum_{\mathrm{x}}\left[\mathrm{A}_0 - I\left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x};\mathbf{p}\right);\mathbf{q}\right)\right)\right]\left[\nabla I\frac{\partial\mathbf{N}}{\partial\mathbf{q}}\right], \tag{4.6b}$$

where

$$\mathrm{H}_1 = \left[\nabla\mathrm{A}_0\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right]^T\left[\nabla\mathrm{A}_0\frac{\partial\mathbf{W}}{\partial\mathbf{p}}\right], \tag{4.7a}$$

$$\mathrm{H}_2 = \left[\nabla I\frac{\partial\mathbf{N}}{\partial\mathbf{q}}\right]^T\left[\nabla I\frac{\partial\mathbf{N}}{\partial\mathbf{q}}\right]. \tag{4.7b}$$

As $\frac{\partial\mathbf{N}}{\partial\mathbf{q}}$ is evaluated at $\mathbf{p} = 0$, H_1 can be precomputed and saved in the memory, while H_2 depends on the current shape and the warped input image gradient, and therefore it should be computed in each iteration. Figure 4.1 shows the steps of the bidirectional warping for inverse compositional algorithm. We call this approach Bi-ICA in the rest of this work.

The “projecting out” technique can be applied to the bidirectional warping, i.e. instead of $\mathrm{SD} = \left[\nabla\mathrm{A}_0\frac{\partial\mathrm{W}}{\partial\mathrm{p}}\right]$ in the Equation (4.6a) and (4.7a), SD is calculated as:

$$\mathrm{SD}(\mathrm{x}) = \nabla\mathrm{A}_0\frac{\partial\mathrm{W}}{\partial\mathrm{p}} - \sum_{i=1}^{m}\left[\sum_{x\in\mathbf{s}_0}\mathrm{A}_i\left(\mathrm{x}\right).\nabla\mathrm{A}_0\frac{\partial\mathrm{W}}{\partial\mathrm{p}}\right]\mathrm{A}_i(\mathrm{x}), \tag{4.8}$$

Similar to the PO, the H_1 can be precomputed, but the dot product of the modified steepest descent images with the error image should be computed in each iteration. The bidirectional warping of the PO is called Bi-PO in the rest of this work.

To have a more generic fitting, we can optimize the shape parameters on the full space of the appearance vectors. In this case, we need to optimize the appearance parameters as well as the shape parameters like the SIC method. The algorithm operates by iteratively

Figure 4.1: The Bidirectional Warping Algorithm

```
Pre-compute:
  (3) Evaluate the gradient $\nabla \mathrm{A}_0$ of the template $\mathrm{A}_0\left(\mathrm{x}\right)$
  (4) Evaluate the Jacobian $\frac{\partial \mathrm{W}}{\partial \mathrm{p}}$ at $\left(\mathrm{x}; 0\right)$
  (5) Compute the steepest descent images $\nabla \mathrm{A}_0 \frac{\partial \mathrm{W}}{\partial \mathrm{p}}$
  (6) Compute the Hessian matrix $\mathrm{H}_1$ using Equation (4.7a)
Iterate:
  (1) Warp $I$ with $\mathrm{W}\left(\mathrm{x};\mathrm{p}\right)$ and $\mathrm{N}\left(\mathrm{x};\mathrm{q}\right)$ to compute
      $I\left(\mathrm{N}\left(\mathrm{W}\left(\mathrm{x;p}\right);\mathrm{q}\right)\right)$
  (2) Compute $E = \left[I\left(\mathrm{N}\left(\mathrm{W}\left(\mathrm{x;p}\right);\mathrm{q}\right)\right) - \mathrm{A}_0\left(\mathrm{x}\right)\right]$
  (7) Evaluate the gradient $\nabla I\left(\mathrm{N}\left(\mathrm{W}\left(\mathrm{x;p}\right);\mathrm{q}\right)\right)$
  (8) Evaluate the Jacobian $\frac{\partial \mathrm{N}}{\partial \mathrm{q}}$
  (9) Compute the steepest descent images $\nabla I \frac{\partial \mathrm{N}}{\partial \mathrm{q}}$
  (10) Compute the Hessian matrix $\mathrm{H}_2$ using Equation (4.7b)
  (11) Compute $\Delta \mathrm{p}$ and $\Delta \mathrm{q}$ using Equation (4.6a) and (4.6b)
  (12) Update $\mathrm{W}\left(\mathrm{x};\mathrm{p}\right) \leftarrow \mathrm{W}\left(\mathrm{x};\mathrm{p}\right) \circ \mathrm{W}\left(\mathrm{x};-\Delta \mathrm{p}\right)^{-1}$
       and $\mathrm{q} = \mathrm{q} + \Delta \mathrm{q}$
```

minimizing:

$$\begin{aligned} f(x) = \sum_{\mathbf{x} \in \mathbf{s}_0} & \left[\mathrm{A}_0\left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x}; 0 + \Delta \mathbf{p}\right); 0\right)\right)\right. \\ & + \sum_{i=1}^{m} \left(\lambda_i + \Delta \lambda_i\right) \mathrm{A}_i \left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x}; 0 + \Delta \mathbf{p}\right); 0\right)\right) \\ & \left. - I\left(\mathbf{N}\left(\mathbf{W}\left(\mathbf{x}; \mathbf{p}\right); \mathbf{q} + \Delta \mathbf{q}\right)\right)\right]^2, \end{aligned} \tag{4.9}$$

simultaneously with respect to $\Delta \mathbf{p}$, $\Delta \mathbf{q}$ and $\Delta \lambda = \left(\Delta \lambda_1, ..., \Delta \lambda_m\right)$. Then we update the warp $\mathbf{W}(\mathbf{x}; \mathbf{p}) \leftarrow \mathbf{W}(\mathbf{x}; \mathbf{p}) \circ \mathbf{W}(\mathbf{x}; \Delta \mathbf{p})^{-1}$, $\mathbf{q} = \mathbf{q} + \Delta \mathbf{q}$ and $\lambda = \lambda + \Delta \lambda$.

We define the concatenation parameter of the shape and the appearance $\mathrm{r} = [\mathrm{p}, \lambda]^T$, and the steepest-descent images as:

$$\mathrm{SD}_{sim}(\mathrm{x}) = \left(\nabla \mathrm{A} \frac{\partial \mathrm{W}}{\partial \mathrm{p}_1}, ..., \nabla \mathrm{A} \frac{\partial \mathrm{W}}{\partial \mathrm{p}_n}, \mathrm{A}_1, ..., \mathrm{A}_m\right) \tag{4.10}$$

where $\nabla \mathrm{A} = \nabla \mathrm{A}_0 + \sum_{i=1}^{m} \lambda_i \nabla \mathrm{A}_i$. We can then compute the parameter update $\Delta \mathrm{r}$ as:

$$\Delta \mathrm{r} = -\mathrm{H}_{sim}^{-1} \sum_{\mathrm{X}} \mathrm{SD}_{sim}^{T}(\mathrm{x}) E(\mathrm{x}) \tag{4.11}$$

where $\mathrm{H}_{sim}^{-1} = \sum_{\mathrm{X}} \mathrm{SD}_{sim}^{T}(\mathrm{x}) \mathrm{SD}_{sim}(\mathrm{x})$.

To find the parameter of the global transformation ($\mathbf{q}$), we used incremental update as: $\mathbf{q} = \mathbf{q} + \Delta \mathbf{q}$, where $\Delta \mathbf{q} = \mathrm{H}_2^{-1} \sum_{\mathrm{X}} -E(\mathrm{x}) \left[\nabla I \frac{\partial \mathbf{N}}{\partial \mathbf{q}} \right]$. This approach is called Bi-SIC in the rest of this work. In this case both SD_{sim} and H_{sim} are calculated in each iteration. The extra computational load of Bi-SIC in comparison with SIC is to calculate the gradient of the warped image and H_2 in each iteration.

In addition to the introduced bidirectional approach, we also propose two modifications to AAM fitting as follows:

1-Affine Transformation: Image alignment techniques for AAM fitting usually consider a 2D set of similarity transform for the global transformation. Affine transformation can improve the performance of Active Shape Model for facial feature extraction (Mahoor et al., 2006). In this work, we apply an affine transformation with six degrees of freedom for AAM fitting. Assuming the base mesh is: $\mathbf{s}_0 = (x_1^0, y_1^0, ..., x_v^0, y_v^0)^{\mathrm{T}}$. We choose $\mathbf{s}_1^* = (x_1^0, 0, ..., x_v^0, 0)^{\mathrm{T}}$, $\mathbf{s}_2^* = (y_1^0, 0, ..., y_v^0, 0)^{\mathrm{T}}$, $\mathbf{s}_3^* = (0, x_1^0, ..., 0, x_v^0)^{\mathrm{T}}$, $\mathbf{s}_4^* = (0, y_1^0, ..., 0, y_v^0)^{\mathrm{T}}$, $\mathbf{s}_5^* = (1, 0, \ldots, 1, 0)^{\mathrm{T}}$ and $\mathbf{s}_6^* = (0, 1, \ldots, 0, 1)^{\mathrm{T}}$. The global affine transformation is defined as: $\mathbf{N}(\mathbf{x}; \mathbf{q}) = \mathbf{s}_0 + \sum_{i=1}^{6} q_i \mathbf{s}_i^*$. This transformation has more degrees of freedom and therefore results in a better modeling of the shape variation.

2- Fitting Constraint: Introduced approaches for AAM fitting still suffer from lack of generality for unseen faces. In addition, the result can differ significantly from trained shapes. One idea is to apply some constraints on fitting iterations. Defining a well constraint is not easy because of the complexity of the face shape, huge variation of the appearance due to different subjects, illuminations and expressions, and the existence of non-face areas (e.g. glasses). In this dissertation, we apply a simple constraint of Active Shape Models (ASM) (Cootes et al., 2004), i.e. those shape parameters (p) are updated that $p_i \leq 3\sqrt{b_i}$, where b_i are the eigenvalues of the trained shapes. This constraint will force the algorithm to result in shapes similar to trained shapes with a limited degree of freedom and therefore prevent it from resulting in non-face shapes.

4.2.3 Experimental Results

We implemented PO (Matthews and Baker, 2004), SIC (Gross et al., 2005), and our proposed Bi-PO and Bi-SIC methods using Matlab platform. We also used the affine transformation for the global transformation instead of 2D similarity and applied the introduced constraint to the PO, SIC, Bi-PO, and Bi-SIC methods and called them PO-AC, SIC-AC, Bi-PO-AC, and Bi-SIC-AC, respectively.

We applied the aforementioned methods on CMU Multi-PIE face dataset (Gross et al., 2010). The CMU Multi-PIE database contains more than 750,000 images of 337 people. Subjects were imaged under 15 view points and 19 illumination conditions. The image resolution is 640×480, where the distance between the center of the eyes are approximately 80 pixels. Certain poses of a subset have 68 facial landmark points. We select a subset from the dataset containing 100 different subjects with the frontal head pose and with the same illumination. We also selected 50 images of left and right head poses that have 68 facial landmark points. Figure 4.2 shows images of sample subjects in frontal, left and right poses.

To initialize the shape model in AAM fitting, we selected two outer eye corners and the chin point (3 points) from the ground truth landmarks and perturbed them randomly by 5 pixels. Then we used the average shape obtained from training subjects as the initial shape and transformed it using similarity transformation obtained by those three perturbed points. Figure 4.3a shows the initial shape for a sample image.

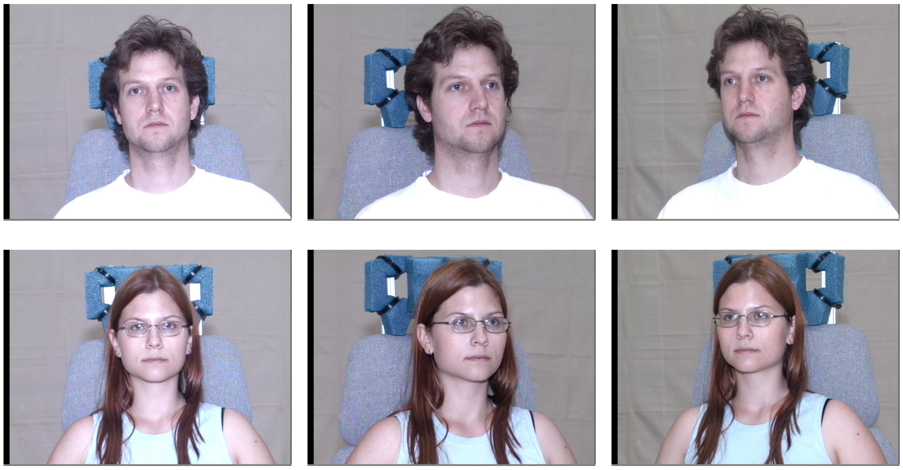
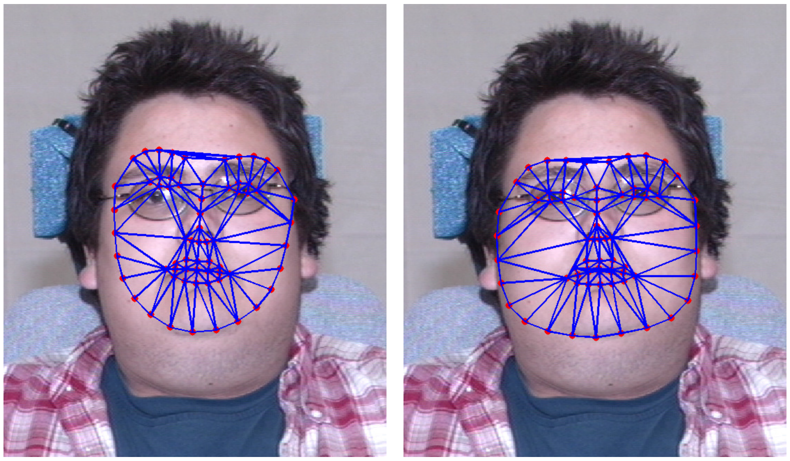


Figure 4.2: Some sample images of frontal, left and right poses from Multi-PIE dataset (Gross et al., 2010).

We tested the performance of the PO, SIC, PO-AC, SIC-AC, Bi-PO, Bi-SIC, Bi-PO-AC, and Bi-SIC-AC methods when the number of images in the training sets varied using 10-fold cross validation. Particularly, we selected 10, 20, 30, 40, 50, 60, 70, 80 and 90 images randomly from the frontal subset and trained separate AAMs. For testing the generalization performance of the fitting methods, we fitted the trained models onto 10 images that are not included in the training sets and repeated this experiment 10 times for different test images. For comparing the fitting performance, we calculated the Root Mean Square Error (RMSE). The value of RMSE shows the distance between the fitted and the actual shape. Naturally, the smaller the RMSE, the better the fitting.

In our first experiment, we examined the effect of using affine transformation and constraint on both the PO and SIC method as well as the introduced bidirectional warping. Figure 4.4 shows the fitting RMSE value of the PO, Bi-PO, PO-AC, and Bi-PO-AC on the frontal subset. Figure 4.5 shows the fitting RMSE value of the SIC, Bi-SIC, SIC-AC, and Bi-SIC-AC on the frontal subset. In both experiments, using affine transformation and having constraint improved the fitting performance. When we have the constraint, it keeps



(a) initial shape (b) fitted shape

Figure 4.3: Initial and fitted shapes of a sample image.

the shape similar to the trained shapes (i.e. face) during the fitting process and prevents the algorithm from resulting non-face shapes. In addition, the affine transformation gives the algorithm more degrees of freedom, and therefore it fits better on unseen samples. It is also shown that bidirectional warping has a better fitting performances than unidirectional warping. Bi-PO and Bi-SIC both have comparative fitting performance and both fit better in comparison with the original unidirectional algorithms.

There are no standard or established choices for the convergence criterion. In this work, we visually inspected a number of results in the RMSE range of 0-20 and confirmed that those having RMSE less than 5 pixels seem successfully fitted. Figure 4.3b shows a sample fitted image having RMSE 4.02.

Figure 4.6 shows the percentage of fitted shapes for the frontal subset using PO, Bi-PO, PO-AC, and Bi-PO-AC. Figure 4.7 shows the percentage of fitted shapes for the frontal subset using SIC, Bi-SIC, SIC-AC and Bi-SIC-AC. As it shown, the bidirectional warping has better performance than the unidirectional method. Also applying the constraint and affine transformation result in a better modeling of unseen images and more convergence on both the PO and SIC. It should be mentioned that the percentage of fitting depends on

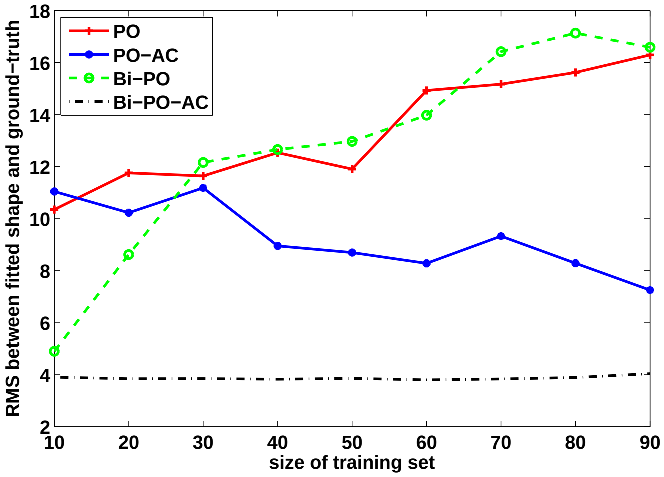


Figure 4.4: RMSE of fitting for variation of Projecting Out as: Projecting Out (PO), Projecting Out with Affine constraint (PO-AC), Bidirectional Projecting Out (Bi-PO), and Bidirectional Projecting Out with Affine constraint (PO-AC).

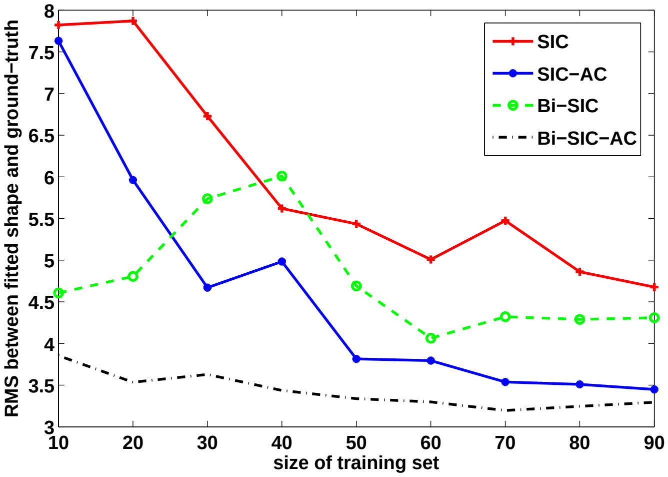


Figure 4.5: RMSE of fitting for variation of SIC. Simultaneously Inverse Compositional (SIC), Simultaneously Inverse Compositional with Affine Constraint (SIC-AC), Bidirectional Simultaneously Inverse Compositional (Bi-SIC), Bidirectional Simultaneously Inverse Compositional with Affine Constraint (Bi-SIC-AC).

the threshold value, but empirically both algorithms have more or less similar performance in comparison to each other in a reasonable range of threshold value.

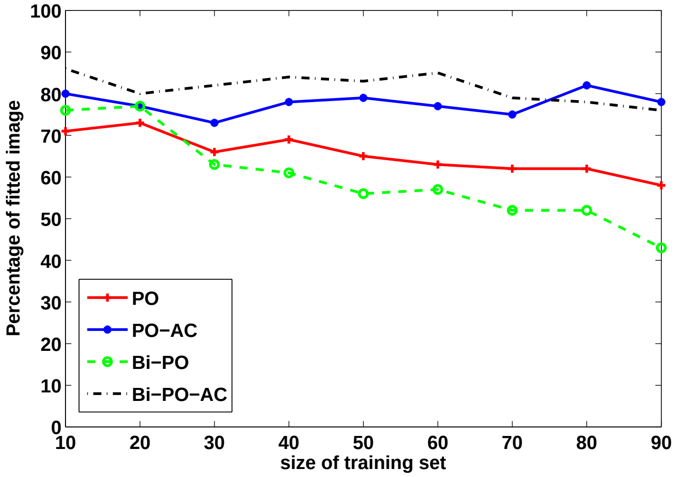


Figure 4.6: Percentage of fitted images for variation of PO.

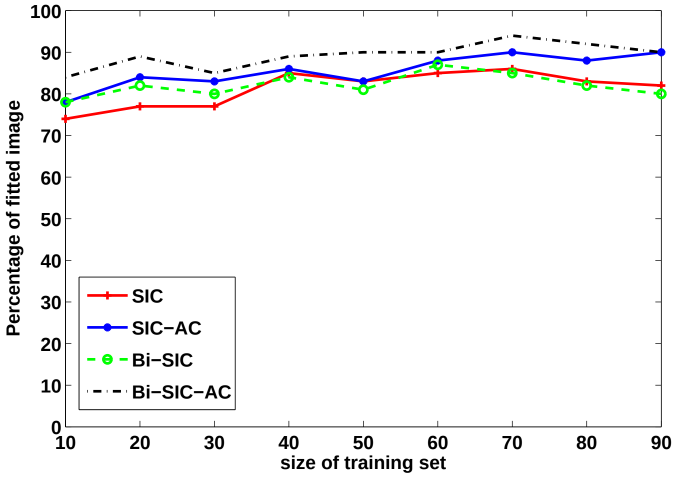


Figure 4.7: Percentage of fitted images for variation of SIC.

In another experiment, we tested the generalization performance of our proposed approach for different poses. We trained an AAM with 120 images (40 images of each frontal, left and right subsets). To test the generality of the fitting, we fitted the trained model onto the 10 other subjects from each pose. We repeated this experiment five times and averaged the fitting results of the SIC, Bi-SIC, SIC-AC, and Bi-SIC-AC. Initial shape was again the warped average shape obtained from training subjects. Table 4.1 shows the average RMSE

Table 4.1: RMSE of fitting on the left and right poses.

	SIC	SIC-AC	Bi-SIC	Bi-SIC-AC
left	6.99	8.60	8.43	8.76
right	4.07	3.40	4.02	3.37
frontal	3.78	3.46	4.02	3.38

Table 4.2: Percentage of fitted on the left and right poses.

	SIC	SIC-AC	Bi-SIC	Bi-SIC-AC
left	72	76	62	72
right	80	88	80	90
frontal	86	90	84	96

of fitting for frontal, left and right poses. Similarly, we defined a threshold of RMSE less than 5 pixels as the fitted shape. Table 4.2 shows the percentage of fitted shapes for frontal, left and right pose subsets. Similar to the previous experiment, using affine transformation and applying constraints on SIC improve the fitting performance. The introduced bidirectional approach also improves the SIC performance significantly, especially when we have pose variations.

Computational Complexity: The bidirectional method introduces an extra computation in every iterations of fitting. If we assume n is the number of warp parameters, N is the number of pixels, and m is the number of top appearance eigenvectors, the complexity of the PO and SIC methods per iteration are $O(nN + n^2)$ and $O((n+m)^2N + (n+m)^3)$, respectively (Baker and Matthews, 2004). In the bidirectional approach, we have k parameters for the chosen global transformation, and in every iterations we need to compute: the gradient of the image (step 7) with the complexity of $O(N)$; the Jacobian $\frac{\partial \mathrm{N}}{\partial \mathrm{q}}$ (step 8) with the complexity of $O(kN)$; the steepest descent images (step 9) with the complexity of $O(kN)$; the Hessian matrix H_2 and invert it (step 10) with the complexity of $O(k^2N + k^3)$; and $\Delta\mathrm{q}$ with complexity of $O(kN + k)$.

The complexity overload of the bidirectional approach is $O(k^2N + k^3)$. The numbers n and m depend on the size of the training set and the model dimensionalities. In most AAM implementations, the dimensionalities of the shape and appearance models are chosen by retaining a fixed percentage (typically 95$\%$) of the variance in the eigenvalues (Gross et al., 2005). In our experimental results, depending on the size of the training set, n varies between $[10, 30]$ and m varies between $[12, 70]$. For the affine transformation, k is 6. Hence, the complexity of Bi-PO is at least two times greater than PO, and the complexity of Bi-SIC is greater than SIC. However, this is based on the assumption of having the same constant factor for all steps.

We implemented all algorithms using Matlab on a windows platform. We executed them on a PC with Intel core Duo 3.00 GHz CPU having 4 GB of RAM, where both implementations have the same termination condition, i.e. the algorithm terminates if the shape does not change or continues for 50 iterations at maximum. In practice, the implemented PO and SIC methods take 3 and 8 seconds for each frame, while the execution of the Bi-PO and Bi-SIC-AC methods take 20 and 27 seconds, respectively.

4.2.4 Discussion and Conclusions

In summary, unlike previous image alignment approaches for AAM fitting that warp either the input image (e.g. Lucas-Kanade method) or the appearance template (e.g. inverse compositional algorithm), we warp both the input image for the global transformation and the template for the shape parameters in the fitting process. Warping both the input image and the appearance template causes the AAM to consider more appearance variations, and therefore it can fit better on images with different poses and appearances. We showed that the introduced bidirectional approach can be applied on the “projected out” and the “simultaneously inverse compositional” approaches for AAM fitting. We also proposed using affine transformation with six degrees of freedom instead of 2D similarity and applying a simple constraint to prevent the fitting algorithm from resulting in shapes far from face geometry.

We tested the performance of the proposed approach on Mutli-PIE dataset. We compared the accuracy of our proposed fitting approach with the PO and SIC methods. First, we trained the AAM with different number of training images and tested the fitting accuracy on unseen images. In another experiment, we then compared the accuracy of fitting on images with different poses. Our experimental results showed that warping both the image and the template makes the AAM fitting more generic. In addition, applying affine transformation gives the algorithm more degrees of freedom to model new face instances and the proposed constraint in the fitting iterations prevents resulting in non-face shapes. In conclusion, our method is promising for modeling and tracking facial images of unseen subjects (i.e. generic model) and also when the accuracy of AAM fitting has priority to the execution speed.

4.3 A New Deep Neural Network Architecture

In this section, we present a novel deep neural network architecture for the FER problem, and examines the network's ability to perform cross-database classification while training on databases that have limited scope, and are often specialized for a few expressions (e.g. MultiPIE and FERA). We conducted comprehensive experiments on seven well-known facial expression databases (viz. MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013) and obtain results which are significantly better than, or comparable to, traditional convolutional neural networks or other state-of-the-art methods in both accuracy and learning time.

Often improving neural network architectures has relied on increasing the number of neurons or increasing the number of layers, allowing the network to learn more complex functions; however, increasing the depth and complexity of a topology leads to a number of problems such as increased over-fitting of training data, and increased computational needs. A natural solution to the problem of increasingly dense networks is to create deep sparse networks, which has both biological inspiration, and has firm theoretical foundations discussed in (Arora et al., 2013). Unfortunately, current GPUs and CPUs do not have the capability to efficiently compute actions on sparse networks. The Inception layer presented in (Russakovsky et al., 2014) attempts to rectify these concerns by providing an approximation of sparse networks to gain the theoretical benefits proposed by Arora et al. (2013), however retains the dense structure required for efficient computation.

Applying the Inception layer to applications of Deep Neural Network has had remarkable results (Sun et al., 2015; Szegedy et al., 2014), and it seems only logical to extend state of the art techniques used in object recognition to the FER problem. In addition to merely providing theoretical gains from the sparsity, and thus, relative depth, of the network, the Inception layer also allows for improved recognition of local features, as smaller convolutions are applied locally, while larger convolutions approximate global features. The increased local performance seems to align logically with the way that humans process emotions as well. By looking at local features such as the eyes and mouth, humans can distinguish the majority of the emotions (Bal et al., 2010). Similarly, children with autism often cannot distinguish emotion properly without being told to remember to look at the same local features (Bal et al., 2010). By using the Inception layer structure and applying the network-in-network theory proposed by Lin et al. (2013b), we can expect significant gains on local feature performance, which seems to logically translate to improved FER results.

Another benefit of the network-in-network method is that along with increased local performance, the global pooling performance is increased and therefore it is less prone to overfitting. This resistance to overfitting allows us to increase the depth of the network significantly without worrying about the small corpus of images that we are working with in the FER problem.

The proposed DNN architecture is inspired by the techniques provided by the GoogLeNet and AlexNet architectures. Our network consists of two elements, first our network contains of two traditional CNN modules (a traditional CNN layer consists of a convolution layer by a max pooling layer). Both of these modules use rectified linear units (ReLU) which have an activation function described by:

$$f(x) = max(0, x) \tag{4.12}$$

where x is the input to the neuron. Using the ReLU activation function allows us to avoid the vanishing gradient problem caused by some other activation functions (Krizhevsky et al., 2012). Following these modules, we apply the techniques of the network in network architecture and add two ”Inception” style modules, which are made up of a 1×1, 3×3 and 5×5 convolution layers (Using ReLU) in parallel. These layers are then concatenated as output and we use two fully connected layers as the classifying layers (Also using ReLU). Figure 4.8 shows the architecture of the network used in this research.

In this work, we register facial images in each of the databases using research standard techniques. We used bidirectional warping of Active Appearance Model (AAM) (Mollahosseini and Mahoor, 2013) and a Supervised Descent Method (SDM) called IntraFace (Xiong and De la Torre, 2013) to extract facial landmarks, however further work could consider improving the landmark recognition in order to extract more accurate faces. IntraFace uses SIFT features for feature mapping and trains a descent method by a linear regression on

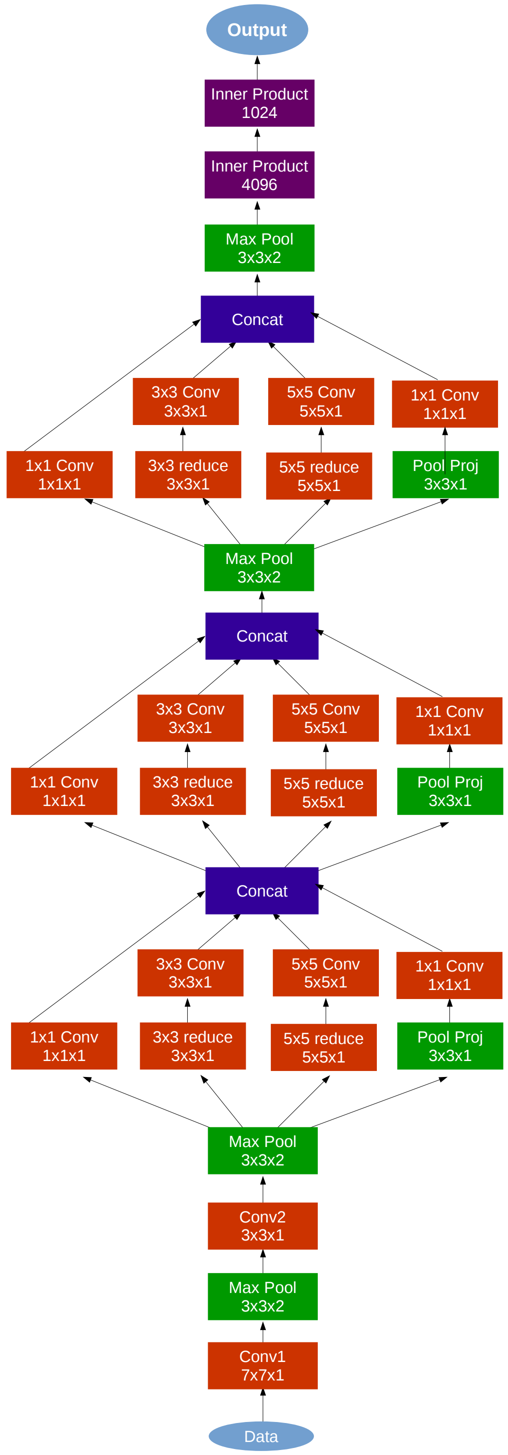


Figure 4.8: Network Architecture

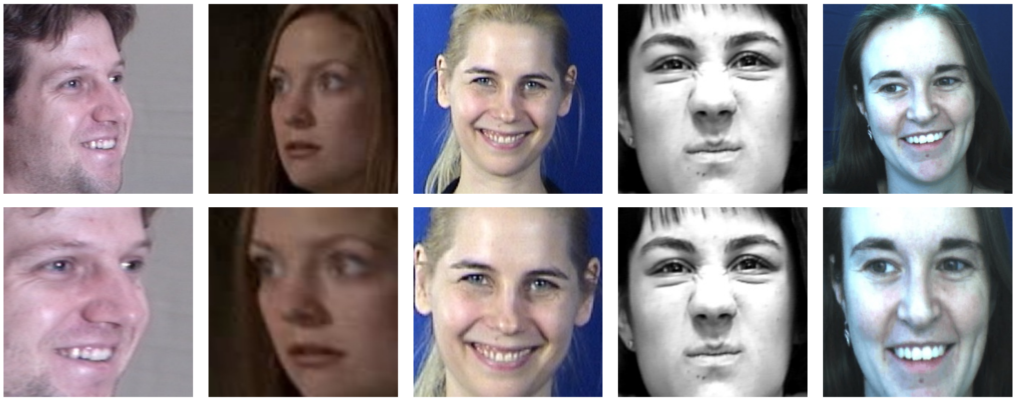


Figure 4.9: Sample of the face registration. From left to right images are taken from MultiPIE, SFEW, MMI, CK+ and DISFA. First row shows the original images and the second row shows their registered images respectively.

training set in order to extract 49 points. We use these points to register faces to an average face in an affine transformation. Finally, a fixed rectangle around the average face is considered as the face region. Figure 4.9 demonstrates samples of the face registration with this method. In our research, facial registration increased the accuracy of our FER algorithms by 4-10%, which suggests that registration (like normalization in traditional problems) is a significant portion of any FER algorithm.

Once the faces have been registered, the images are resized to 48×48 pixels for analysis. Even though many databases are composed of images with a much higher resolution testing suggested that decreasing this resolution does not greatly impact the accuracy, however vastly increases the speed of the network. To augment our data, we extract 5 crops of 40×40 from the four corners and the center of the image and utilize both of them and their horizontal flips for a total of 10 additional images.

In training the network, the learning rates are decreased in a polynomial fashion as: $base\_lr(1 - iter/max\_iter)^{0.5}$, where $base\_lr = 0.01$ is the base learning rate, $iter$ is the current iteration and max_iter is the maximum allowed iterations. Testing suggested that other popular learning rate policies such as *fixed* learning rate, *step* where learning rate is

Table 4.3: Proposed DNN architecture Configuration

Layer type	**Patch Size/ Stride**	**Output**	**1 x 1**	**3 x 3**	**3 x 3 reduce**	**5 x 5**	**5 x 5 reduce**	**Pooling**	**Operations**
Convolution - 1	7×7 / 2	$24 \times 24 \times 64$							5.7M
Max pool - 1	3×3 / 2	$12 \times 12 \times 64$							5.7M
Convolution - 2	3×3 / 1	$12 \times 12 \times 192$							1.4M
Max Pool - 2	3×3 / 2	$6 \times 6 \times 192$							1.4M
Inception - 3a			64	128	96	32	16	32	2.6M
Inception - 3b			128	192	128	96	32	64	4.5M
Max Pool - 4	3×3 / 2	$3 \times 3 \times 480$							0.6M
Inception - 4a			192	208	96	48	16	64	1.3M
Avg Pooling - 6		$1 \times 1 \times 1024$							25.6K
Fully Connected		$1 \times 1 \times 4096$							0.2M
Fully Connected		$1 \times 1 \times 1024$							51K

multiplies by a gamma factor in each step, and exponential approach did not perform as well as the polynomial fashion. Using the polynomial learning rate, the test loss converged faster and allowed us to train the network for many iterations without the need for fine-tuning. We also trained the bias nodes twice as fast as the weights of the network, in order to increase the rate at which unnecessary nodes are removed from evaluation. This decreases the number of iterations that the network must run before the loss converges.

4.3.1 Face Databases

In the FER problem, however, unlike visual object databases such as imageNet (Deng et al., 2009), existing FER databases often have limited numbers of subjects, few sample images or videos per expression, or small variation between sets, making neural networks significantly more difficult to train. For example, the FER2013 database (Goodfellow et al., 2015) (one of the largest recently released FER databases) contains 35,887 images of different subjects yet only 547 of the images portray disgust. Similarly, the CMU MultiPIE face database (Gross et al., 2010) contains around 750,000 images but it is comprised of

Table 4.4: Number of images per each expression in databases

	AN	DI	FE	HA	NE	SA	SU
MultiPie	0	22696	0	47338	114305	0	19817
MMI	1959	1517	1313	2785	0	2169	1746
CK+	45	59	25	69	0	28	83
DISFA	436	5326	4073	28404	48582	1024	1365
FERA	1681	0	1467	1882	0	2115	0
SFEW	104	81	90	112	98	92	86
FER2013	4953	547	5121	8989	6198	6077	4002

* AN, DI, FE, HA, Ne, SA, SU stand for Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprised respectively.

only 337 different subjects, where 348,000 images portray only a "neutral" emotion and the remaining images do not portray anger, fear or sadness.

Since, many samples are necessary for a DNN to extract most appropriate and distinguishable features, we evaluate the proposed method on well-known publicly available facial expressions databases: CMU MultiPIE (Gross et al., 2010), MMI (Pantic et al., 2005), Denver Intensity of Spontaneous Facial Actions (DISFA) (Mavadati et al., 2013), extended CK+ (Lucey et al., 2010), GEMEP-FERA database (Bänziger and Scherer, 2010), SFEW (Dhall et al., 2011), and FER2013 (Goodfellow et al., 2015).

Table 4.4 shows the number of images for six basic expressions and neutral faces in each database.

4.3.2 Results

We evaluated the accuracy of the proposed deep neural network architecture in two different experiments; viz. subject-independent and cross-database evaluation. In the subject-independent experiment, databases are split into training, validation, and test sets in a strict subject independent manner. We used the K-fold cross validation technique with K=5 to evaluate the results. In FERA and SFEW, the training and test sets are defined in the

Table 4.5: Average Accuracy (%) for subject-independent

	Top-1	**Top-2**	**State-of-the-arts**
MultiPIE	94.7±0.8	98.7±0.3	70.6 (Lee et al., 2014), 90.6 (Eleftheriadis et al., 2015)
MMI	77.6±2.9	86.8±6.2	63.4 (Liu et al., 2014a), 74.7 (Liu et al., 2013), 79.8 (Mayer et al., 2014), 86.9 (Shan et al., 2009)
DISFA	55.0±6.8	69.8±8.6	-
FERA	76.7±3.6	90.5±4.6	56.1 (Liu et al., 2014a), 75.0 (UCR-team, 2011), 55.6 (Valstar et al., 2011)
SFEW	47.7±1.7	62.1±1.2	26.1 (Liu et al., 2013), 24.7 (Eleftheriadis et al., 2015)
CK+	93.2±1.4	97.8±1.3	84.1 (Mayer et al., 2014), 84.4 (Lee et al., 2014), 88.5 (Taheri et al., 2014), 92.0 (Liu et al., 2013) 92.4 (Liu et al., 2014a), 93.6 (Zhang et al., 2015)
FER2013	66.4±0.6	81.7±0.3	69.3Tang (2013)

database release, and the results are evaluated on the database defined test set without performing K-fold cross validation. Since there are different samples per emotion per subject in some databases, the training, validation and test sets have slightly different sample sizes in each fold. On average we used 175K samples for training, 56K samples for validation, and 64K samples for test. The proposed architecture was trained for 200 epochs (i.e. 150K iterations on mini-batches of size 250 samples). Table 4.5 gives the average accuracy when classifying the images into the six basic expressions and the neutral expression. The average confusion matrix for subject-independent experiments can be seen in Table 4.6.

Here, we also report the top-2 expression classes. As Table 4.5 depicts, the accuracy of the top-2 classification is 15% higher than the top-1 accuracy in most cases, especially in the wild datasets (i.e. FERA, SFEW, FER2013). We believe that by assigning a single expression to a image can be ambiguous when there is transition between expressions or the given expression is not at its peak, and therefore the top-2 expression can result in a better classification performance when evaluating image sequences.

The proposed architecture was implemented using the Caffe toolbox (Jia et al., 2014) on a Tesla K40 GPU with 2880 CUDA cores and 12GB RAM which is able to perform 4.29 TFLOPS single precision operation. It took roughly 20 hours to train 175K samples for 200

Table 4.6: Average (%) confusion matrix for subject-independent

		predicted						
		AN	**DI**	**FE**	**HA**	**NE**	**SA**	**SU**
Actual	**AN***	**55.0**	7.0	12.8	3.5	7.6	8.5	5.3
	DI	1.0	**80.3**	1.8	5.8	8.5	2.2	0.1
	FE	7.4	4.3	**47.0**	8.1	18.7	8.6	5.5
	HA	0.7	3.2	2.4	**86.6**	5.5	0.2	1.0
	NE	2.3	6.3	7.8	5.5	**75.0**	1.3	1.4
	SA	6.0	11.3	8.9	2.7	13.7	**56.1**	0.9
	SU	0.8	0.1	2.8	3.5	2.5	0.6	**89.3**

* AN, DI, FE, HA, Ne, SA, SU stand for Anger, Disgust, Fear, Happiness, Neutral, Sadness, Surprised respectively.

epochs. Figure 4.10 shows the training loss and classification accuracy of the top-1 and top-2 classification labels on the validation set of the subject-independent experiment over 150,000 iterations (about 150 epochs). As the figure illustrates, the proposed architecture converges after about 50 epochs.

In the cross-database experiment, one database is used for evaluation and the rest of databases are used to train the network. Because every database has a unique fingerprint (lighting, pose, emotions, etc.) the cross database task is much more difficult to extract features from (both for traditional SVM approaches, and for neural networks). The proposed architecture was trained for 100 epochs in each experiment. Table 4.7 gives the average cross-database accuracy when classifying the six basic expressions as well as the neutral expression.

As a benchmark to our proposed solution, we trained a full AlexNet from scratch (as opposed to fine tuning an already trained network) using the same protocol as used to train our own network. As shown in Table 4.8, our proposed architecture has better performance on MMI & FER2013 and comparable performance on the rest of the databases. The value of the proposed solution over the AlexNet architecture is its training time - Our version of

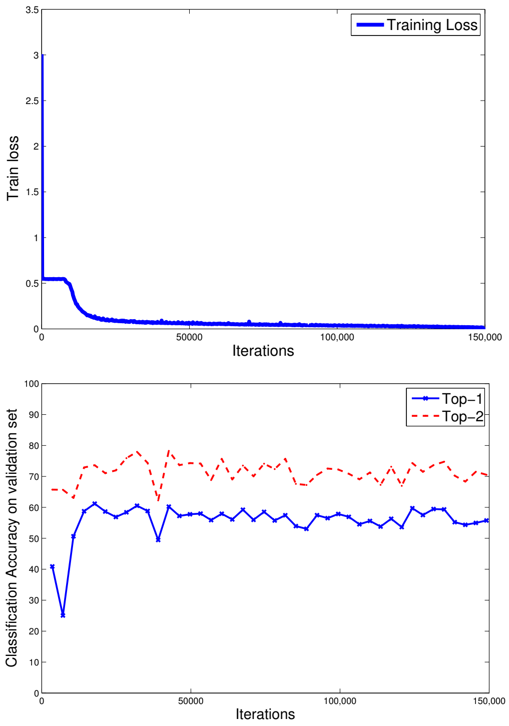


Figure 4.10: Training loss and classification accuracy on validation set

Table 4.7: Average Accuracy (%) on cross database

	Top-1	**Top-2**	**Mayer et al. (2014)**	**Shan et al. (2009)**	**Miao et al. (2012)**	**Zhang et al. (2015)**
MultiPIE	45.7	63.2	-	-	-	-
MMI	55.6	68.3	51.4	50.8	36.8	66.9
DISFA	37.7	53.2	-	-	-	-
FERA	39.4	58.7	-	-	-	-
SFEW	39.8	55.3	-	-	-	-
CK+	64.2	83.1	47.1	-	56.0	61.2
FER2013	34.0	51.7	-	-	-	-

Table 4.8: Subject-independent comparison with AlexNet results (% accuracy)

	Proposed Architecture	**AlexNet**
MultiPie	94.7	94.8
MMI	77.9	56.0
DISFA	55.0	56.1
FERA	76.7	77.4
SFEW	47.7	48.6
CK+	93.2	92.2
FER2013	66.4	61.1

AlexNet performed more than 100M operations, whereas the proposed network performs about 25M operations.

4.3.3 Discussion

As shown in Tables 4.5 and 4.7, the results in the subject-independent tests were either comparable to or better than the current state of the art. It should be mentioned that we have compared our results with the best methods on each database separately, where the hyper parameters of the presented models are fine-tuned for that specific problem. We perform significantly better than the state of the art on MultiPIE and SFEW (no known state of the art has been reported for the DISFA database). The only exceptions to the improved performance are with the MMI and FERA databases. There are a number of explanations for this phenomenon.

It should be mentioned that, we have compared our results with the best methods on each dataset separately, where hyper parameters are fine-tuned for that specific problem. For example, Mayer et al. (2014) has a better accuracy on MMI dataset compared to ours (79.8% to 77.6%), while the same method lower accuracy on CK+ dataset, or Shan et al. (2009) performs better than our proposed approach in the case of subject-independent, while we got a better result in cross-database experiments.

One of the likely reasons for the performance discrepancies on the subject-independent databases is due to the way that the networks are trained in our experiments. Because we use data from all of the studied databases to train the deep architecture, the input data contains image that do not conform to the database setting such as pose and lighting. It is very difficult to avoid this issue as it is hard or impossible to train such a complex network architecture on so little data without causing significant overfitting. Another reason for the decreased performance is the focus on cross-database performance. By training slightly less complicated architectures, or even using traditional methods such as support vector machines, or engineered features, it would likely be possible to improve the performance of the network on subject-independent tasks. In this research however, we present a comprehensive solution that can generalize well to the FER "in the wild" problem.

4.4 AffectNet

There are several models in the literature to quantify affective facial behaviors: 1) categorical model, where the emotion/affect is chosen from a list of affective-related categories such as six basic emotions defined by Ekman and Friesen (1971), 2) dimensional model, where a value is chosen over a continuous emotional scale, such as valence and arousal (Russell, 1980) and 3) Facial Action Coding System (FACS) model, where all possible facial actions are described in terms of Action Units (AUs) (Ekman and Friesen, 1977). FACS model explains facial movements and does not describe the affective state directly. There are several methods to convert AUs to affect space (e.g., EMFACS (Friesen and Ekman, 1983) states that the occurrence of AU6 and AU12 is a sign of happiness). In the categorical model, mixed emotions cannot adequately be transcribed into a limited set of words. Some researchers tried to define multiple distinct compound emotion categories (e.g., happily surprised, sadly fearful) (Du et al., 2014) to overcome this limitation. However, still the set is limited, and the intensity of the emotion cannot be defined in the categorical model. In contrast, the dimensional model of affect can distinguish between subtly different displays of affect and encode small changes in the intensity of each emotion on a continuous scale, such as valence and arousal. Valence refers to how positive or negative an event is, and arousal reflects whether an event is exciting/agitating or calm/soothing (Russell, 1980). Figure 4.11 shows samples of facial expressions represented in the 2D space of valence and arousal. As it is shown, there are several different kinds of affect and small changes in the same emotion that cannot be easily mapped into a limited set of terms existing in the categorical model.

There is a debate that the reduction of emotion space into two dimensions may not cover the full affect space. Fontaine et al. (2007) suggested four continuous dimensions to represent similarities and differences in the meaning of emotion words. Although having more dimensions can cover a larger affect space, still valence and arousal are the most accepted and commonly used continuous dimensions in the affective computing community.

Recently, databases of facial expression and affect in the wild received much attention. These databases are either captured from movies or the Internet, and annotated with categorical model (Dhall et al., 2013; Goodfellow et al., 2015; Mollahosseini et al., 2016b), dimensional model (Zafeiriou et al., 2016), and FACS model (Benitez-Quiroz et al., 2016). However, they only cover one model of affect, have a limited number of subjects, or contain few samples of certain emotions such as disgust. Therefore, a large database, with a large amount of subject variations in the wild condition that covers multiple models of affect (especially the dimensional model) is a need.

To address this need, we created a database of facial **Affect** from the Inter**Net** (called *AffectNet*) by querying different search engines (Google, Bing, and Yahoo) using 1250 emotion related tags in six different languages (English, Spanish, Portuguese, German, Arabic, and Farsi). *AffectNet* contains more than one million images with faces and extracted facial

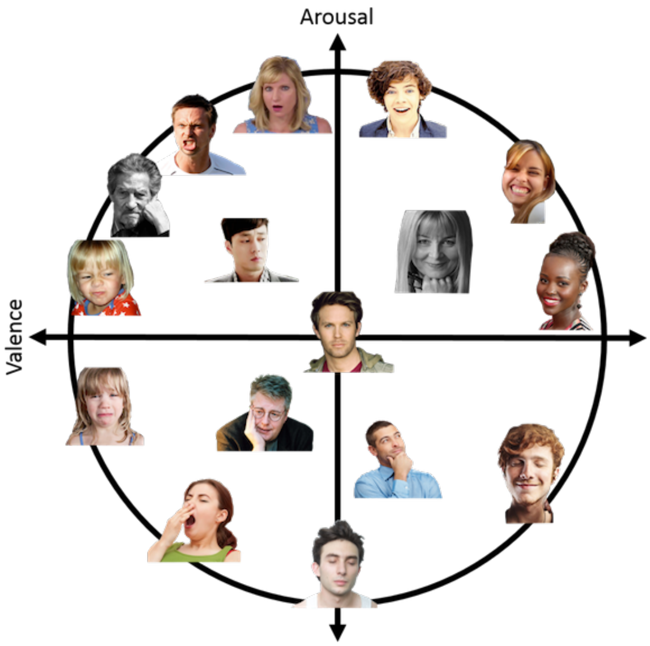


Figure 4.11: Sample images in Valence Arousal circumplex

landmark points. Twelve human experts manually annotated 440,000 of these images in both categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face. Figure 4.11 shows sample images from *AffectNet* and their valence and arousal annotations.

To calculate the agreement level between the human labelers, 36,000 images were annotated by two human labelers. *AffectNet* is by far the largest database of facial affect in still images which covers both categorical and dimensional models. The URL of facial images, their facial landmark points, the query terms, and the affect labels will be publicly available to the research community [1].

[1] A copy of AffectNet is available in: http://mohammadmahoor.com/databases-codes/

4.4.1 Existing databases

Early databases of facial expressions such as JAFFE (Lyons et al., 1998), XM2VTS (Messer et al., 1999), Cohn-Kanade (Lucey et al., 2010; Tian et al., 2001), MMI (Pantic et al., 2005), and MultiPie (Gross et al., 2010) were captured in a lab-controlled environment where the subjects portrayed different facial expressions. This approach resulted in a clean and high-quality database of posed facial expressions. However, posed expressions may differ from daily life unposed (aka spontaneous) facial expressions. Thus, capturing spontaneous expression became a trend in the affective computing community. Examples of these environments are recording the responses of participants' faces while watching a stimuli (e.g., DISFA (Mavadati et al., 2013), AM-FED (McDuff et al., 2013)) or performing laboratory-based emotion inducing tasks (e.g., Belfast (Sneddon et al., 2012)). These databases often capture multi-modal affects such as voice, biological signals, etc. and usually a series of frames are captured that enable researchers to work on temporal and dynamic aspects of expressions. However, the diversity of these databases is limited due to the number of subjects, head pose variation, and environmental conditions.

Hence there is a demand to develop systems that are based on natural, unposed facial expressions. To address this demand, recently researchers paid attention to databases in the wild. Dhall et al. (2013) released *Acted Facial Expressions in the Wild (AFEW)* from 54 movies by a recommender system based on subtitles. The video clips were annotated with six basic expressions plus neutral. AFEW contains 330 subjects aged 1-77 years and addresses the issue of temporal facial expressions in the wild. A static subset (SFEW (Dhall et al., 2011)) is created by selecting some frames of AFEW. SFEW covers unconstrained facial expressions, different head poses, age range, occlusions, and close to real world illuminations. However, it contains only 700 images, and there are only 95 subjects in the database.

The *Facial Expression Recognition 2013 (FER-2013)* database was introduced in the ICML 2013 Challenges in Representation Learning (Goodfellow et al., 2015). The database was created using the Google image search API that matched a set of 184 emotion-related keywords to capture the six basic expressions as well as the neutral expression. Images were resized to 48x48 pixels and converted to grayscale. Human labelers rejected incorrectly labeled images, corrected the cropping if necessary, and filtered out some duplicate images. The resulting database contains 35,887 images most of which are in the wild settings. FER-2013 is currently the biggest publicly available facial expression database in the wild settings, enabling many researchers to train machine learning methods such as Deep Neural Networks (DNNs) where large amounts of data are needed. In FER-2013, the faces are not registered, a small number of images portray disgust (547 images), and unfortunately most of facial landmark detectors fail to extract facial landmarks at this resolution and quality. In addition, only the categorical model of affect is provided with FER-2013.

The *Affectiva-MIT Facial Expression Dataset (AM-FED)* database (McDuff et al., 2013) contains 242 facial videos (160K frames) of people watching Super Bowl commercials using their webcam. The recording conditions were arbitrary with different illumination and contrast. The database was annotated frame-by-frame for the presence of 14 FACS action units, head movements, and automatically detected landmark points. AM-FED is a great resource to learn AUs in the wild. However, there is not a huge variance in head pose (limited profiles), and there are only a few subjects in the database.

The FER-Wild (Mollahosseini et al., 2016b) database contains 24,000 images that are obtained by querying emotion-related terms from three search engines. The OpenCV face recognition was used to detect faces in the images, and 66 landmark points were found using Active Appearance Model (AAM) (Mollahosseini and Mahoor, 2013) and a face alignment algorithm via regression local binary features (Ren et al., 2014; Yu, 2016). Two human labelers annotated the images into six basic expressions and neutral. Comparing with FER-2013, FER-Wild images have a higher resolution with facial landmark points necessary to register the images. However, still a few samples portray some expressions such as disgust and fear and only the categorical model of affect is provided with FER-Wild.

The *EmotioNet* (Benitez-Quiroz et al., 2016) consists of one million images of facial expressions downloaded from the Internet by selecting all the words derived from the word “feeling” in WordNet (Miller, 1995). Face detector (Viola and Jones, 2004) was used to detect faces in these images and the authors visually inspected the resultant images. These images were then automatically annotated with AUs and AU intensities by an approach based on Kernel Subclass Discriminant Analysis (KSDA) (You et al., 2011). The KSDA-based approach was trained with Gabor features centered on facial landmark with a Radial Basis Function (RBF) kernel. Images were labeled as one of the 23 (basic or compound) emotion categories defined in (Du et al., 2014) based on AUs. For example, if an image has been annotated as having AUs 1, 2, 12 and 25, it is labeled as happily surprised. A total of 100,000 images (10% of the database) were manually annotated with AUs by experienced coders. The proposed AU detection approach was trained on CK+ (Lucey et al., 2010), DISFA (Mavadati et al., 2013), and CFEE (Lucey et al., 2011) databases, and the accuracy of the automated annotated AUs was reported about 80% on the manually annotated set. EmotioNet is a novel resource of FACS model in the wild with a large amount of subject variation. However, it lacks the dimensional model of affect, and the emotion categories are defined based on annotated AUs and not manually labeled.

On the other hand, some researchers developed databases of the dimensional model in the continuous domain. These databases, however, are limited since the annotation of continuous dimensions is more expensive and necessitate trained annotators. Examples of these databases are *Belfast* (Sneddon et al., 2012), *RECOLA* (Ringeval et al., 2013), *Affectiva-MIT Facial Expression Dataset (AM-FED)* (McDuff et al., 2013), and recently published *Aff-Wild Database* (Zafeiriou et al., 2016) which is the only database of dimensional model in the wild.

The Belfast database (Sneddon et al., 2012) contains recordings (5s to 60s in length) of mild to moderate emotional responses of 60 participants to a series of laboratory-based emotion inducing tasks (e.g., surprise response by setting off a loud noise when the participant is asked to find something in a black box). The recordings were labeled by information on self-report of emotion, the gender of the participant/experimenter, and the valence in the continuous domain. The arousal dimension was not annotated in Belfast database. While the portrayed emotions are natural and spontaneous, the tasks have taken place in a relatively artificial setting of a laboratory where there was a control on lighting conditions, head poses, etc.

The *Database for Emotion Analysis using Physiological Signals (DEAP)* (Koelstra et al., 2012) consists of spontaneous reactions of 32 participants in response to one-minute long music video clip. The EEG, peripheral physiological signals, and frontal face videos of participants were recorded, and the participants rated each video in terms of valence, arousal, like/dislike, dominance, and familiarity. Correlations between the EEG signal frequencies and the participants' ratings were investigated, and three different modalities, i.e., EEG signals, peripheral physiological signals, and multimedia features on video clips (such as lighting key, color variance, etc.) were used for binary classification of low/high arousal, valence, and liking. DEAP is a great database to study the relation of biological signals and dimensional affect, however, it has only a few subjects and the videos are captured in lab controlled settings.

The RECOLA benchmark (Ringeval et al., 2013) contains videos of 23 dyadic teams (46 participants) that participated in a video conference completing a task which required collaboration. Different multi-modal data of the first five minutes of interaction, i.e., audio, video, ECG and EDA) were recorded continuously and synchronously. Six annotators measured arousal and valence. The participants reported their arousal and valence through the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994) questionnaire before and after the task. RECOLA is a great database of the dimensional model with multiple cues and modalities, however, it contains only 46 subjects and the videos were captured in the lab controlled settings.

Audio-Visual Emotion recognition Challenge (AVEC) series of competitions (Ringeval et al., 2015; Schuller et al., 2011, 2012; Valstar et al., 2013, 2014, 2016) provided a benchmark of automatic audio, video and audiovisual emotion analysis in continuous affect recognition. AVEC 2011, 2012, 2013, and 2014 used videos from the SEMAINE (McKeown et al., 2012) database videos. Each video is annotated by a single rater for every dimension using a two-axis joystick. AVEC 2015 and 2016 used the RECOLA benchmark in their competitions. Various continuous affect recognition dimensions were explored in each challenge year such as valence, arousal, expectation, power, and dominance, where the prediction of valence and arousal are studied in all challenges.

The *Aff-Wild* Database (Zafeiriou et al., 2016) is by far the largest database for measuring continuous affect in the valence-arousal space “in-the-wild”. More than 500 videos from YouTube were collected. Subjects in the videos displayed a number of spontaneous emotions while watching a particular video, performing an activity, and reacting to a practical joke. The videos have been annotated frame-by-frame by three human raters, utilizing a joystick-based tool to rate valence and arousal. *Aff-Wild* is a great database of dimensional modeling in the wild that considers the temporal changes of the affect, however, it has a small subject variance, i.e., it only contains 500 subjects.

Table 4.9 summarizes the characteristics of the reviewed databases in all three models of affect, i.e., categorical model, dimensional model, and Facial Action Coding System (FACS).

Table 4.9: The Summary and Characteristics of Reviewed Databases in Affect Recognition

Database	Database information	# of Subjects	Condition	Affect Modeling
CK+ Lucey et al. (2010)	- Frontal and 30 degree images	- 123	- Controlled - Posed	- 30 AUs - 7 emotion categories
MultiPie Gross et al. (2010)	- Around 750,000 images - Under multiple viewpoints and illuminations	- 337	- Controlled - Posed	- 7 emotion categories
MMI Pantic et al. (2005)	- Subjects portrayed 79 series of facial expressions - Image sequence of frontal and side view are captured	- 25	- Controlled - Posed & Spontaneous	- 31 AUs - Six basic expression
DISFA Mavadati et al. (2013)	- Video of subjects while watching a four minutes video - Clip are recorded by a stereo camera	- 27	- Controlled - Spontaneous	- 12 AUs
SALDB Nicolaou et al. (2010) Nicolaou et al. (2011)	- SAL - Audiovisual (facial expression,shoulder, audiocues) - 20 facial feature points, 5 shoulder points for video	- 4	- Controlled - Spontaneous	- Valence - Quantized - Continuous
RELOCA Ringeval et al. (2013)	- Multi-modal audio, video, ECG and EDA	- 46	- Controlled - Spontaneous	- Valence and arousal (continuous) - Self assessment
AM-FED McDuff et al. (2013)	- 242 facial videos	- 242	- Spontaneous	- 14 AUs
DEAP Koelstra et al. (2012)	- 40 one-minute long videos shown to subjects - EEG signals recorded	- 32	- Controlled - Spontaneous	- Valence and arousal - Self assessment
AFEW Dhall et al. (2013)	- Videos	- 330	- Wild	- 7 emotion categories
FER-2013 Goodfellow et al. (2015)	- Images queried from web	- $\sim$35,887	- Wild	- 7 emotion categories
EmotioNet Benitez-Quiroz et al. (2016)	- Images queried from web - 100,000 images annotated manually - 900,000 images annotated automatically	- $\sim$100,000	- Wild	- 12 AUs annotated - 23 emotion categories based on AUs
Aff-Wild Zafeiriou et al. (2016)	- 500 videos from YouTube	- 500	- Wild	- Valence and arousal (continuous)
FER-Wild Mollahosseini et al. (2016b)	- 24,000 images from web	- $\sim$24,000	- Wild	- 7 emotion categories
AffectNet **(This work)**	**- 1,000,000 images with facial landmarks** **- 440,000 images annotated manually**	**- $\sim$440,000**	**- Wild**	**- 8 emotion categories** **- Valence and arousal**

4.4.2 Evaluation Metrics

There are various evaluation metrics in the literature to measure the reliability of annotation and automated affective computing systems. Accuracy, F1-score (Sokolova et al., 2006), Cohen's kappa (Cohen, 1960), Krippendorf's Alpha (Krippendorff, 1970), ICC (Shrout and Fleiss, 1979), area under the ROC curve (AUC), and area under Precision-Recall curve (AUC-PR) (Jeni et al., 2013) are well-defined widely used metrics for evaluation of the categorical and FACS-based models. Since, the dimensional model of affect is usually evaluated in a continuous domain, different evaluation metrics are necessary. In the following, we review several metrics that are used in the literature for evaluation of dimensional model.

Root Mean Square Error (RMSE) is the most common evaluation metric in a continuous domain which is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2} \tag{4.13}$$

where $\hat{\theta}_i$ and θ_i are the prediction and the ground truth of i^{th} sample, and n is the number of samples in the evaluation set. RMSE-based evaluation can heavily weigh the outliers (Bermejo and Cabestany, 2001), and it is not able to provide the covariance of prediction and ground-truth to show how they change with respect to each other. Pearson's correlation coefficient is therefore proposed in some literature (Nicolaou et al., 2011; Schuller et al., 2011, 2012) to overcome this limitation:

$$CC = \frac{COV\{\hat{\theta}, \theta\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E[(\hat{\theta} - \mu_{\hat{\theta}})(\theta - \mu_{\theta})]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{4.14}$$

Concordance Correlation Coefficient (CCC) is another metric (Ringeval et al., 2015; Valstar et al., 2016) which combines the Pearson's correlation coefficient (CC) with the square difference between the means of two compared time series:

$$\rho_c = \frac{2\rho\sigma_{\hat{\theta}}\sigma_{\theta}}{\sigma_{\hat{\theta}}^2 + \sigma_{\theta}^2 + (\mu_{\hat{\theta}} - \mu_{\theta})^2} \tag{4.15}$$

where ρ is the Pearson correlation coefficient (CC) between two time-series (e.g., prediction and ground-truth), $\sigma_{\hat{\theta}}^2$ and σ_{θ}^2 are the variance of each time series, and $\mu_{\hat{\theta}}$ and μ_{θ} are the mean value of each. Unlike CC, the predictions that are well correlated with the ground-truth but shifted in value are penalized in proportion to the deviation in CCC.

The value of valence and arousal are [-1,+1] and their signs are essential in many emotion-prediction applications. For example, if the ground-truth valence is +0.3, prediction of +0.7 is far better than prediction of -0.1, since +0.7 indicates a positive emotion similar to the ground-truth (despite both predictions have the same RMSE). Sign Agreement Metric (SAGR) is another metric that is proposed in (Nicolaou et al., 2011) to evaluate the performance of a valence and arousal prediction system. SAGR is defined as:

$$SAGR = \frac{1}{n}\sum_{i=1}^{n} \delta(sign(\hat{\theta}_i), sign(\theta_i)) \tag{4.16}$$

where δ is the Kronecker delta function, defined as:

$$\delta(a,b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{4.17}$$

The above discussed metrics are used to evaluate the categorical and dimensional baselines on *AffectNet* in Sec. 4.5.

4.4.3 Existing Algorithms

Affective computing is now a well-established field, and there are many algorithms and databases for developing automated affect perception systems. Since it is not possible to include all those great works, we only give a brief overview and cover the state-of-the-art methods that are applied on the databases explained in Sec. 4.4.1.

Conventional algorithms of affective computing from faces use hand-crafted features such as facial landmarks (Kobayashi and Hara, 1997), pixel intensities (Mohammadi et al., 2014), Gabor filters (Liu and Wechsler, 2002), Local Binary Patterns (LBP) (Shan et al., 2009), Local Phase Quantization (LPQ) (Zhen and Zilu, 2012), and Histogram of Oriented Gradients (HOG) (Mavadati et al., 2013). These hand-crafted features often lack enough generalizability in the wild settings where there is a high variation in scene lighting, camera view, image resolution, background, subjects head pose and ethnicity.

An alternative approach is to use Deep Neural Networks (DNN) to learn the most appropriate feature abstractions directly from the data and handle the limitations of hand-crafted features. DNNs have been a recent successful approach in visual object recognition (Krizhevsky et al., 2012), human pose estimation (Toshev and Szegedy, 2014), face verification (Taigman et al., 2014) and many more. This success is mainly due to the availability of computing power and existing big databases that allow DNNs to extract highly discriminative features from the data samples. There have been enormous attempts on using DNNs in automated facial expression recognition and affective computing (Fan et al., 2016; He et al., 2015; Mollahosseini et al., 2016a,b; Tang, 2013) that are especially very successful in the wild settings.

Table 4.10 shows a list of the state-of-the-art algorithms and their performance on the databases listed in Table 4.9. As shown in the table, the majority of these approaches have used DNNs to learn a better representation of affect, especially in the wild settings. Even some of the approaches, such as the winner of the AVEC 2015 challenge (He et al., 2015), trained a DNN with hand-crafted features and still could improve the prediction accuracy.

Table 4.10: State-of-the-art Algorithms and Their Performance on the Databases Listed in Table 4.9.

Work	Database	Method	Results
Mollahosseini et al. (2016a)	CK+ MultiPie	- Inception based Convolutional Neural Network (CNN) - Subject-independent and cross-database experiments	- 93.2% accuracy on CK+ - 94.7% accuracy on MultiPie
Shan et al. (2009)	MMI	- Different SVM kernels trained with LBP features - Subject-independent and cross-database experiments	- 86.9% accuracy on MMI
Zhang et al. (2015)	DISFA	- l_pnorm multi-task multiple kernel learning - learning shared kernels from a given set of base kernels	- 0.70 F1-score on DISFA - 0.93 recognition rate on DISFA
Nicolaou et al. (2011)	SALDB	- Bidirectional LSTM - Trained on multiple engineered features extracted from audio, facial geometry , and shoulder	- Leave-one-sequence-out - BLSTM-NN outperform SVR - Valence (RMSE=0.15 and CC=0.796) - Arousal (RMSE=0.21 and CC=0.642)
He et al. (2015)	RECOLA	- Multiple stack of bidirectional LSTM (DBLSTM-RNN) - Trained on engineered features extracted from audio, video (LPQ-TOP), 52 ECG features, and 22 EDA features	- Winner of AVEC 2015 challenge - Valence (RMSE=0.104 and CC=0.616) - Arousal (RMSE=0.121 and CC=0.753)
McDuff et al. (2013)	AM-FED	- HOG features extracted - SVM with RBF kernel	- AUC 0.90, 0.72 and 0.70 for smile, AU2 and AU4 respectively
Koelstra et al. (2012)	DEAP	- Gaussian naive Bayes classifier - EEG, physiological signals, and multimedia features - Binary classification of low/high arousal and valence,	- 0.39 F1-score on Arousal - 0.37 F1-score on Valence - 0.40 F1-score on Liking
Fan et al. (2016)	AFEW	- Trained on both video and audio. - VGG network are followed by LSTMs and combined with 3D convolution	- Winner of EmotiW 2016 challenge - 56.16% accuracy on AFEW
Tang (2013)	FER-2013	- CNN with linear one-vs-all SVM at the top	- Winner of the FER challenge - 71.2% accuracy on test set
Benitez-Quiroz et al. (2016)	EmotioNet	- New face feature extraction method using Gabor filters - KSDA classification - Subject-independent and cross-database experiments	- ∼80% AU detection on EmotioNet
Mollahosseini et al. (2016b)	FER-Wild	- Trained on AlexNet - Noise estimation methods used	- 82.12% accuracy on FER-Wild

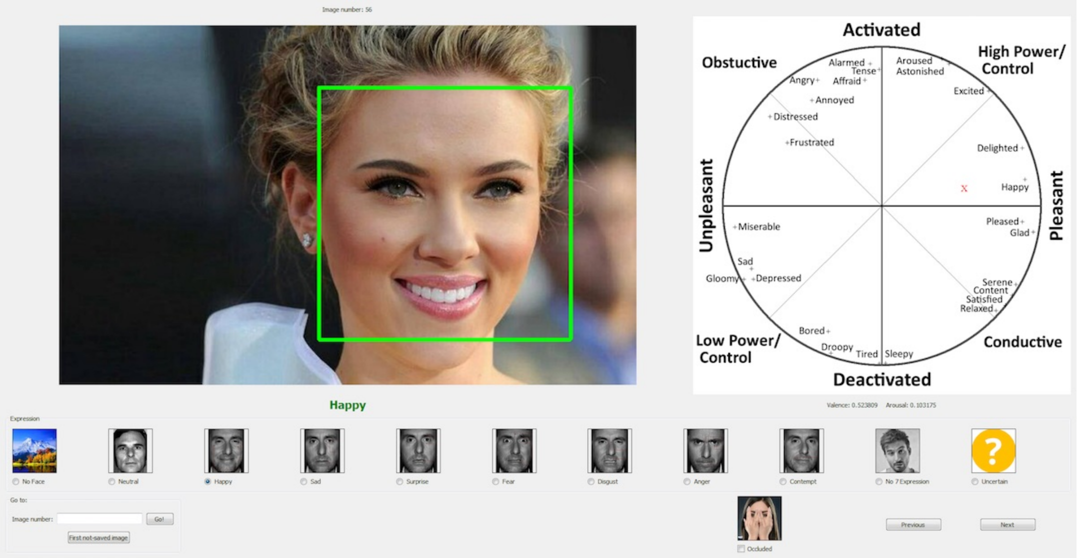


Image number: 56

Happy

Valence: 0.523809 Arousal: 0.103175

Expression

○ No Face ○ Neutral ● Happy ○ Sad ○ Surprise ○ Fear ○ Disgust ○ Anger ○ Contempt ○ No 7 Expression ○ Uncertain

Go to:

Image number: [Go!]

[First not-saved image]

□ Occluded

[Previous] [Next]

Figure 4.12: A screen-shot of the software application used to annotate categorical and dimensional (valence and arousal) models of affect and the osculation tag if existing. Only one detected face in each image is annotated (shown in the green bounding box).

4.4.4 Facial Images from the Web

Emotion-related keywords were combined with words related to gender, age, or ethnicity, to obtain nearly 362 strings in the English language such as “joyful girl”, “blissful Spanish man”, “furious young lady”, “astonished senior”. These keywords are then translated into five other languages: Spanish, Portuguese, German, Arabic and Farsi. The direct translation of queries in English to other languages did not accurately result in the intended emotions since each language and culture has differing words and expressions for different emotions. Therefore, the list of English queries was provided to native non-English speakers who were proficient in English, and they created a list of queries for each emotion in their native language and inspected the quality of the results visually. The criteria for high-quality queries were those that returned a high percentage of human faces showing the intended queried emotions rather than drawings, graphics, or non-human objects. A total of 1250 search queries were compiled and used to crawl the search engines in our database.

Since a high percentage of results returned by our query terms already contained neutral facial images, no individual query was performed to obtain additional neutral face.

Three search engines (Google, Bing, and Yahoo) were queried with these 1250 emotion related tags. Other search engines such as Baidu and Yandex were considered. However, they either did not produce a large number of facial images with intended expressions or they did not have available APIs for automatically querying and pulling image URLs into the database. Additionally, queries were combined with negative terms (e.g., “drawing”, “cartoon”, “animation”, “birthday”, etc.) to avoid non-human objects as much as possible. Furthermore, since the images of stock photo websites are posed unnaturally and contain watermarks mostly, a list of popular stock photo websites was compiled and the results returned from the stock photo websites were filtered out.

A total of $\sim$1,800,000 distinct URLs returned for each query were stored in the database. The OpenCV face recognition was used to obtain bounding boxes around each face. A face alignment algorithm via regression local binary features (Ren et al., 2014; Yu, 2016) was used to extract 66 facial landmark points. The facial landmark localization technique was trained using the annotations provided from the 300W competition (Sagonas et al., 2013, 2016). More than 1M images containing at least one face with extracted facial landmark points were kept for further processing.

4.4.5 Annotation

Crowd-sourcing services like Amazon Mechanical Turk are fast, cheap and easy approaches for labeling large databases. The quality of labels obtained from crowd-sourcing services, however, varies considerably among the annotators. Due to these issues and the fact that annotating the valence and arousal requires a deep understanding of the concept, we avoided crowd-sourcing facilities and instead hired 12 full-time and part-time annotators at the University of Denver to label the database. A total of 440,000 images were given to these expert annotators to label the face in the images into both discrete categorical and continuous dimensional (valence and arousal) models. Due to time and budget constraints each image was annotated by one annotator.

A software application was developed to annotate the categorical and dimensional (valence and arousal) models of affect. Figure 4.12 shows a screen-shot of the annotation application. A comprehensive tutorial including the definition of the categorical and dimensional models of affect with some examples of each category, valence and arousal was given to the annotators. Three training sessions were provided to each annotator, in which the annotator labeled the emotion category, valence and arousal of 200 images and the results were reviewed with the annotators. Necessary feedback was given on both the categorical and dimensional labels. In addition, the annotators tagged the images that have any occlusion on the face. The occlusion criterion was defined as if any part of the face was not visible. If the person in the images wore glasses, but the eyes were visible without any shadow, it was not considered as occlusion.

Categorical Model Annotation

Eleven discrete categories were defined in the categorical model of *AffectNet* as: Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face. The *None* ("None of the eight emotions") category is the type of expression/emotions (such as sleepy, bored, tired, seducing, confuse, shame, focused, etc.) that could not be assigned by annotators to any of the six basic emotions, contempt or neutral. However, valence and arousal could be assigned to these images. The *Non-face* category was defined as images that: 1) Do not contain a face in the image; 2) Contain a watermark on the face; 3) The face detection algorithm fails and the bounding box is not around the face; 4) The face is a drawing, animation, or painted; and 5) The face is distorted beyond a natural or

Table 4.11: Number of Annotated Images in Each Category

Expression	**Number**
Neutral	80,276
Happy	146,198
Sad	29,487
Surprise	16,288
Fear	8,191
Disgust	5,264
Anger	28,130
Contempt	5,135
None	35,322
Uncertain	13,163
Non-Face	88,895

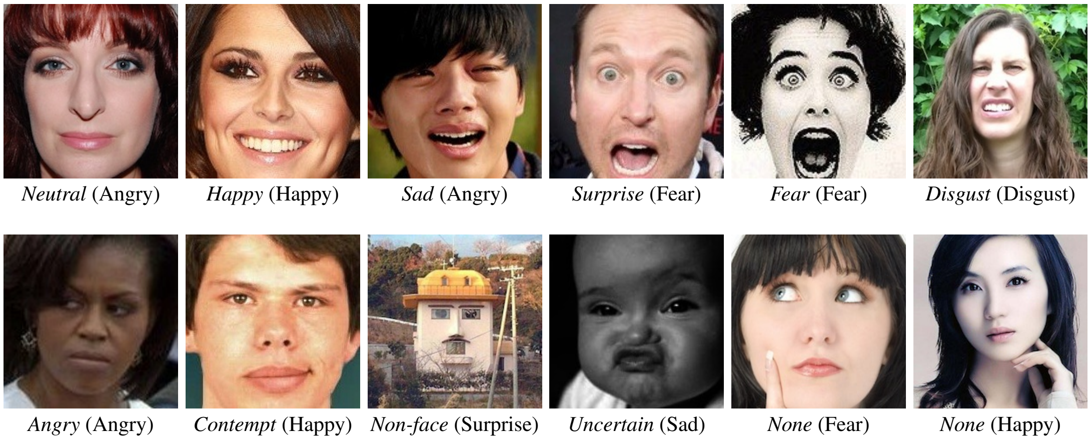
normal shape, even if an expression could be inferred. If the annotators were uncertain about any of the facial expressions, images were tagged as *uncertain*. When an image was annotated as *Non-face* or *uncertain*, valence and arousal were not assigned to the image.

The annotators were instructed to select the proper expression category of the face, where the intensity is not important as long as the face depicts the intended emotion. Table 4.11 shows the number of images in each category. Table 4.12 indicates the percentage of annotated categories for queried emotion terms. As shown, the happy emotion had the highest hit-rate (48%), and the rest of the emotions had hit-rates less than 20%. About 15% of all query results were in the *No-Face* category, as many images from the web contain watermarks, drawings, etc. About 15% of all queried emotions resulted in neutral faces. Among other expressions, disgust, fear, and contempt had the lowest hit-rate with only 2.7%, 4%, and 2.4% hit-rates, respectively. As one can see, the majority of the returned images from the search engines were happy or neutral faces. The authors believe that this is because people tend to publish their images with positive expressions rather than negative expressions. Figure 4.13 shows a sample image in each category and its intended queries (in parentheses).

Table 4.12: Percentage of Annotated Categories for Queried Emotion Terms (%)

		Query Expression						
		HA	SA	SU	FE	DI	AN	CO
Annotated Expression	NE*	17.3	16.3	13.9	17.8	17.8	16.1	20.1
	HA	**48.9**	**27.2**	**30.4**	**28.6**	**33**	**29.5**	**30.1**
	SA	2.6	15.7	4.8	5.8	4.5	5.4	4.6
	SU	2.7	3.1	16	4.4	3.6	3.4	4.1
	FE	0.7	1.2	4.2	4	1.5	1.4	1.3
	DI	0.6	0.7	0.7	0.9	2.7	1.1	1
	AN	2.8	4.5	3.8	5.6	6	12.2	6.1
	CO	1.3	0.9	0.4	1.1	1.1	1.2	2.4
	NO	5.4	8.7	4.8	8.1	8.8	9.3	11.2
	UN	1.3	3.1	4.3	3.1	4.1	3.7	2.7
	NF	16.3	18.6	16.7	20.6	16.9	16.8	16.3

* NE, HA, SA, SU, FE, DI, AN, CO, NO, UN , and NF stand for Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face categories, respectively.



Neutral (Angry) *Happy* (Happy) *Sad* (Angry) *Surprise* (Fear) *Fear* (Fear) *Disgust* (Disgust)

Angry (Angry) *Contempt* (Happy) *Non-face* (Surprise) *Uncertain* (Sad) *None* (Fear) *None* (Happy)

Figure 4.13: Samples of queried images from the web and their annotated tags. The queried expression is written in parentheses.

Dimensional (Valence & Arousal) Annotation

The definition of valence and arousal dimensions was adapted from Russell (1980) and was given to annotators in our tutorial as: "Valence refers to how positive or negative an event is, and arousal reflects whether an event is exciting/agitating or calm/soothing". A sample circumplex with estimated positions of several expressions, borrowed from Paltoglou and Thelwall (2013), was provided in the tutorial as a reference for the annotators. The provided circumplex in the tutorial contained more than 34 complex emotions categories such as suspicious, insulted, impressed, etc., and used to train annotators. The annotators were instructed to consider the intensity of valence and arousal during the annotation. During the annotation process, the annotators were supervised closely and constant necessary feedback was provided when they were uncertain about some images.

To model the dimensional affect of valence and arousal, a 2D Cartesian coordinate system was used where the x-axis and y-axis represent the valence and arousal, respectively. Similar to Russell's circumplex space model Russell (1980), our annotation software did not allow the value of valence and arousal outside of the circumplex. This allows us to convert the Cartesian coordinates to polar coordinates with $0 \leq r \leq 1$ and $0 \leq \theta < 360$. The annotation software showed the value of valence and arousal to the annotators when they selected a point in the circumplex. This helped the annotators to pick more precise locations of valence and arousal with a higher confidence.

A predefined estimated region of valence and arousal was defined for each categorical emotion in the annotation software (e.g., for happy emotion the valence is in (0.0, 1.0], and the arousal is in [-0.2, 0.5]). If the annotators select a value of valence and arousal outside of the selected emotion's region, the software indicates a warning message. The annotators were able to proceed, and they were instructed to do so, if they were confident about the value of valence and arousal. The images with the warning messages were marked in the

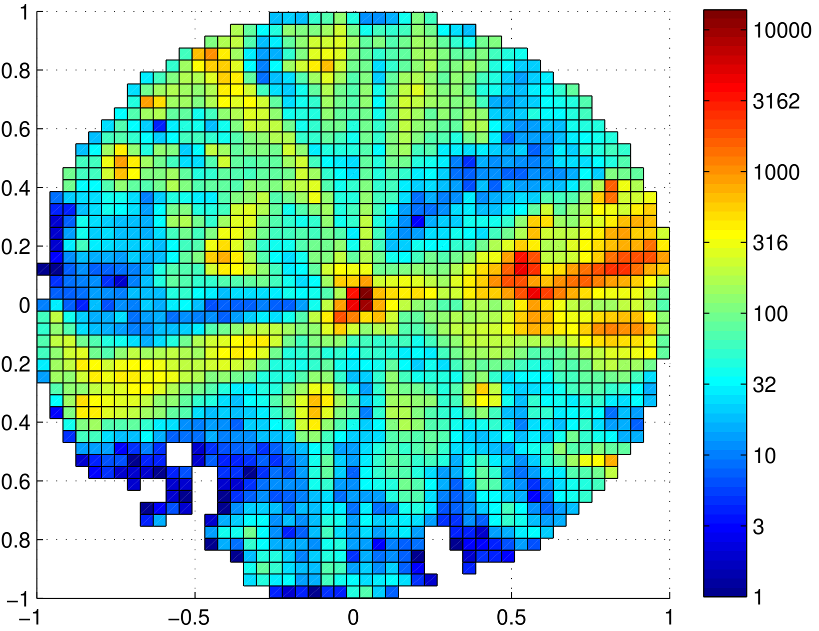


Figure 4.14: Histogram (number of frames in each range/area) of valence and arousal annotations (Best viewed in color).

database, for further review by the authors. This helped to avoid mistakes in the annotation of the dimensional model of affect.

Figure 4.14 shows the histogram (number of samples in each range/area) of annotated images in a 2D Cartesian coordinate system. As illustrated, there are more samples in the center and the right middle (positive valence and small positive arousal) of the circumplex, which confirms the higher number of Neutral and Happy images in the database compared to other categories in the categorical model. [2]

[2] A numerical representation of annotated images in each range/area of valence and arousal is provided in the Appendix B.

Table 4.13: Annotators' Agreement in Dimensional Model of Affect

	Same Category		All	
	Valence	Arousal	Valence	Arousal
RMSE *	0.190	0.261	0.340	0.362
CORR	0.951	0.766	0.823	0.567
SAGR	0.906	0.709	0.815	0.667
CCC	0.951	0.746	0.821	0.551

* RMSE, CORR, SAGR, and CCC stand for Root Mean Square Error, Correlation, Sign Agreement Metric, and Concordance Correlation Coefficient respectively.

4.4.6 Annotation Agreement

In order to measure the agreement between the annotators, 36,000 images were annotated by two annotators. The annotations were performed fully blind and independently, i.e., the annotators were not aware of the intended query or other annotator's response. The results showed that the annotators agreed on 60.7% of the images. Table 4.14 shows the agreement between two annotators for different categories. As it is shown, the annotators highly agreed on the *Happy* and *No Face* categories, and the highest disagreement occurred in the *None* category. Visually inspecting some of the images in the *None* category, the authors believe that the images in this category contain very subtle emotions and they can be easily confused with other categories (the last two example of Fig. 4.13 show images in the *None* category).

Table 4.14: Agreement Between Two Annotators in Categorical Model of Affect (%)

	Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	None	Uncertain	Non-Face
Neutral	**50.8**	7.0	9.1	2.8	1.1	1.0	4.8	5.3	11.1	1.9	5.1
Happy	6.3	**79.6**	0.6	1.7	0.3	0.4	0.5	3.0	4.6	1.0	2.2
Sad	11.8	0.9	**69.7**	1.2	3.4	1.3	4.0	0.3	3.5	1.2	2.6
Surprise	2.0	3.8	1.6	**66.5**	14.0	0.8	1.9	0.6	4.2	1.9	2.7
Fear	3.1	1.5	3.8	15.3	**61.1**	2.5	7.2	0.0	1.9	0.4	3.3
Disgust	1.5	0.8	3.6	1.2	3.5	**67.6**	13.1	1.7	2.7	2.3	2.1
Anger	8.1	1.2	7.5	1.7	2.9	4.4	**62.3**	1.3	5.5	1.9	3.3
Contempt	10.2	7.5	2.1	0.5	0.5	4.4	2.1	**66.9**	3.7	1.5	0.6
None	**22.6**	12.0	14.5	8.0	6.0	2.3	16.9	1.3	9.6	4.3	2.6
Uncertain	13.5	12.1	7.8	7.3	4.0	4.5	6.2	2.6	12.3	**20.6**	8.9
Non-Face	3.7	3.8	1.7	1.1	0.9	0.4	1.7	0.4	1.2	1.4	**83.9**

Table 4.13 shows various evaluation metrics between the two annotators in the continuous dimensional model of affect. These metrics are defined in Sec. 4.4.2. We calculated these metrics in two scenarios: 1) the annotators agreed on the category of the image; 2) on all images that are annotated by two annotators. As Table 4.13 shows, when the annotators agreed on the category of the image, the annotations have a high correlation and sign agreement (SAGR). According to Table 4.14, this occurred on only 60.7% images. However, there is less correlation and SAGR on overall images, since the annotators had a different perception of emotions expressed in the images. It can also be seen that the annotators agreed on valence more than arousal. The authors believe that this is because the perception of valence (how positive or negative the emotion is) is easier and less subjective than arousal (how excited or calm the subject is) especially in still images. Comparing the metrics in the existing dimensional databases (shown in Table 4.10) with the agreement of human labelers on *AffectNet*, suggest that *AffectNet* is a very challenging database and even human annotations have more RMSE than automated methods on existing databases.

The quality of search engines were also evaluated between the indented emotion of the query and the categorical model annotation. The search engines have some overlaps. The greatest overlap is between Bing and Yahoo, where 40% of the images are in common, compared to Bing and Google (have less than 5% common images) and Yahoo and Google (3% in common). When we queried the engines, Google, Yahoo, and Bing produced $\sim$620,000, $\sim$115,000, and $\sim$150,000 images, respectively out of $\sim$1M images. The reason for having more images from Google in the database is that the Google API allowed us to query different image size/resolutions for the same query. According to the annotated images, the overall accuracies of emotion related queries were 12.8%, 21.7%, and 16.60% for Google, Yahoo, and Bing, respectively.

4.5 Baseline

In this section, two baselines are proposed to classify images in the categorical model and predict the value of valence and arousal in the continuous domain of dimensional model. Since deep Convolutional Neural Networks (CNNs) have been a successful approach to learn appropriate feature abstractions directly from the image and there are many samples in *AffectNet* necessary to train CNNs, we proposed two simple CNN baselines for both categorical and dimensional models. We also compared the proposed baselines with conventional approaches (Support Vector Machines (Cortes and Vapnik, 1995) and Support Vector Regressions (Smola and Vapnik, 1997)) learned from hand-crafted features (HOG). In the following sections, we first introduce our training, validation and test sets, and then show the performance of each proposed baselines.

4.5.1 Test, Validation, and Training Sets

Test set: The subset of the annotated images that are annotated by two annotators is reserved for the test set. To determine the value of valence and arousal in the test set, since there are two responses for one image in the continuous domain, one of the annotations is picked randomly. To select the category of image in the categorical model, if there was a disagreement, a favor was given to the intended query, i.e., if one of the annotators labeled the image as the intended query, the image was labeled with the intended query in the test set. This happened in 29.5% of the images with disagreement between the annotators. On the rest of the images with disagreement, one of the annotations was assigned to the image randomly. Since the test set is a random sampling of all images, it is heavily imbalanced. In other words, there are more than 11,000 images with happy expression while it contains only 1,000 images with contemptuous expression.

Validation set: Five hundred samples of each category is selected randomly as a validation set. The validation set is used for hyper-parameter tuning, and since it is balanced, there is no need for any skew normalization.

Training set: The rest of images are considered as training examples.

4.5.2 Categorical Model Baseline

Facial expression data is usually highly skewed. This form of imbalance is commonly referred to as *intrinsic* variation, i.e., it is a direct result of the nature of expressions in the real world. This happens in both the categorical and dimensional models of affect. For instance, Caridakis et al. (2008) reported that a bias toward quadrant 1 (positive arousal, positive valence) exists in the SAL database. The problem of learning from imbalanced data sets has two challenges. First, training data with an imbalanced distribution often causes learning algorithms to perform poorly on the minority class (He and Garcia, 2009). Second, the imbalance in the test/validation data distribution can affect the performance metrics dramatically. Jeni et al. (2013) studied the influence of skew on imbalanced validation set. The study showed that with exception of area under the ROC curve (AUC), all other studied evaluation metrics, i.e., Accuracy, F1-score, Cohen's kappa (Cohen, 1960), Krippendorf's Alpha (Krippendorff, 1970), and area under Precision-Recall curve (AUC-PR) are affected by skewed distributions dramatically. While AUC is unaffected by skew, precision-recall curves suggested that AUC may mask poor performance. To avoid or minimize skew-biased estimates of performance, the study suggested to report both skew-normalized scores and the original evaluation.

We used AlexNet (Krizhevsky et al., 2012) architecture as our deep CNN baseline. AlexNet consists of five convolution layers, followed by max-pooling and normalization layers, and three fully-connected layers. To train our baseline with an imbalanced training set, four approaches are studied in this research as *Imbalanced learning*, *Down-Sampling*, *Up-Sampling*, and *Weighted-Loss*. The imbalanced learning approach was trained with the imbalanced training set without any change in the skew of the dataset. To train the down-sampling approach, we selected a maximum of 15,000 samples from each class. Since there are less than 15,000 samples for some classes such as Disgust, Contempt, and Fear, the resulting training set is semi-balanced. To train the up-sampling approach, we heavily up-sampled the under-represented classes by replicating their samples so that all classes had the same number of samples as the class with maximum samples, i.e., Happy class.

The weighted-loss approach weighted the loss function for each of the classes by their relative proportion in the training dataset. In other words, the loss function heavily penalizes the networks for misclassifying examples from under-represented classes, while penalizing networks less for misclassifying examples from well-represented classes. The entropy loss formulation for a training example (X, l) is defined as:

$$E = -\sum_{i=1}^{K} H_{l,i} log(\hat{p_i}) \tag{4.18}$$

where $H_{l,i}$ denotes row l penalization factor of class i, K is the number of classes, and $\hat{p_i}$ is the predictive softmax with values $[0, 1]$ indicating the predicted probability of each class as:

$$\hat{p_i} = \frac{exp(x_i)}{\sum_{j=1}^{K} exp(x_j)} \tag{4.19}$$

Equation (4.18) can be re-written as:

$$\begin{aligned} E = & -\sum_{i} H_{l,i} log(\frac{exp(x_i)}{\sum_j exp(x_j)}) \\ & = -\sum_{i} H_{l,i} x_i + \sum_{i} H_{l,i} log(\sum_{j} exp(x_j)) \\ & = log(\sum_{j} exp(x_j)) \sum_{i} H_{l,i} - \sum_{i} H_{l,i} x_i \end{aligned} \tag{4.20}$$

The derivate with respect to the prediction x_k is:

$$\begin{aligned} \frac{\partial E}{\partial x_k} &= \frac{\partial}{\partial x_k}[log(\sum_j exp(x_j)) \sum_i H_{l,i}] - \frac{\partial}{\partial x_k}[\sum_i H_{l,i} x_i] \\ &= (\sum_i H_{l,i}) \frac{1}{\sum_j exp(x_j)} \frac{\partial}{\partial x_k} \sum_j exp(x_j) - H_{l,k} \\ &= (\sum_i H_{l,i}) \frac{exp(x_k)}{\sum_j exp(x_j)} - H_{l,k} \\ &= (\sum_i H_{l,i}) \hat{p_k} - H_{l,k} \end{aligned} \tag{4.21}$$

When $H = I$, the identity, the proposed weighted-loss approach gives the traditional cross-entropy loss function. We used the implemented Infogain loss in Caffe (Jia et al., 2014) for this purpose. For simplicity, we used a diagonal matrix defined as:

$$H_{ij} = \begin{cases} \frac{f_i}{f_{min}}, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \tag{4.22}$$

where f_i is the number of samples of the i^{th} class and f_{min} is the number of samples in the most under-represented class, i.e., Disgust class in this situation.

Before training the network, the faces were cropped and resized to 256×256 pixels. No facial registration was performed at this baseline. To augment the data, five crops of 224×224 and their horizontal flips were extracted from the four corners and the center of the image at random during the training phase. The networks were trained for 20 epochs using a batch size of 256. The base learning rate was set to 0.01, and decreased step-wise by a factor of 0.1 every 10,000 iterations. We used a momentum of 0.9.

Table 4.15 shows the top-1 and top-2 F1-Scores for the imbalanced learning, down-sampling, up-sampling, and weighted-loss approaches on the test set. Since the test set is imbalanced, both the skew-normalized and the original scores are reported. The skew normalization is performed by random under-sampling of the classes in the test set. This process is repeated 200 times, and the skew-normalized score is the average of the score on multiple trials. As it is shown, the weighted-loss approach performed better than other approaches in the skew-normalized fashion. The improvement is significant in under-represented classes, i.e., Contempt, Fear, and Disgust. The imbalanced approach performed worst in the Contempt and Disgust categories since there were a few training samples of these classes compared with other classes. The up-sampling approach also did not classify the Contempt and Disgust categories well, since the training samples of these classes were heavily up-sampled (almost 20 times), and the network was over-fitted to these samples. Hence the network lost its generalization and performed poorly on these classes of the test set.

Table 4.16 shows accuracy, F1-score, Cohen's kappa, Krippendorf's Alpha, area under the ROC curve (AUC), and area under the Precision-Recall curve (AUC-PR) on the test sets. Except for the accuracy, all the metrics are calculated in a binary-class manner where the positive class contains the samples labeled by the given category, and the negative class contains the rest. The reported result in Table 4.16 is the average of these metrics over eight classes. The accuracy is defined in a multi-class manner in which the number of correct predictions is divided by the total number of samples in the test set. The skew-normalization is performed by balancing the distribution of classes in the test set using random under-sampling and averaging over 200 trials. Since the validation set is balanced, there is no need for skew-normalization.

Table 4.15: F1-Scores of four different approaches of training AlexNet

	Imbalanced				Down-Sampling				Up-Sampling				Weighted-Loss			
	Top-1		Top-2		Top-1		Top-2		Top-1		Top-2		Top-1		Top-2	
	Orig*	Norm*	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm
Neutral	0.63	0.49	0.82	0.66	0.58	0.49	0.78	0.70	0.61	0.50	0.81	0.64	0.57	0.52	0.81	0.77
Happy	0.88	0.65	0.95	0.80	0.85	0.68	0.92	0.85	0.85	0.71	0.95	0.80	0.82	0.73	0.92	0.88
Sad	0.63	0.60	0.84	0.81	0.64	0.60	0.81	0.78	0.6	0.57	0.81	0.77	0.63	0.61	0.83	0.81
Surprise	0.61	0.64	0.84	0.86	0.53	0.63	0.75	0.83	0.57	0.66	0.80	0.81	0.51	0.63	0.77	0.86
Fear	0.52	0.54	0.78	0.79	0.54	0.57	0.80	0.82	0.56	0.58	0.75	0.76	0.56	0.66	0.79	0.86
Disgust	0.52	0.55	0.76	0.78	0.53	0.64	0.74	0.81	0.53	0.59	0.70	0.72	0.48	0.66	0.69	0.83
Anger	0.65	0.59	0.83	0.80	0.62	0.60	0.79	0.78	0.63	0.59	0.81	0.77	0.60	0.60	0.81	0.81
Contempt	0.08	0.08	0.49	0.49	0.22	0.32	0.60	0.70	0.15	0.18	0.42	0.42	0.27	0.59	0.58	0.79

*Orig and Norm stand for **Orig**inal and skew-**Norma**lized, respectively.

Table 4.16: Evaluation Metrics and Comparison of CNN baselines, SVM and MS Cognitive on Categorical Model of Affect.

	CNN Baselines								**SVM**		**MS Cognitive**	
	Imbalanced		**Down-Sampling**		**Up-Sampling**		**Weighted-Loss**					
	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm	Orig	Norm
Accuracy	0.72	0.54	0.68	0.58	0.68	0.57	0.64	0.63	0.60	0.37	0.68	0.48
F_1-Score	0.57	0.52	0.56	0.57	0.56	0.55	0.55	0.62	0.37	0.31	0.51	0.45
Kappa	0.53	0.46	0.51	0.51	0.52	0.49	0.5	0.57	0.32	0.25	0.46	0.40
Alpha	0.52	0.45	0.51	0.51	0.51	0.48	0.5	0.57	0.31	0.22	0.46	0.37
AUC	0.85	0.80	0.82	0.85	0.82	0.84	0.86	0.86	-	-	0.83	0.77
AUCPR	0.56	0.55	0.54	0.57	0.55	0.56	0.58	0.64	-	-	0.52	0.50

The confusion matrix of the weighted-loss approaches is shown in Table 4.17. The weighted-loss approach classified the samples of Contempt and Disgust categories with an acceptable accuracy but did not perform well in Happy and Neutral. This is because the network was not penalized enough for misclassifying examples from these classes. We believe that a better formulation of the weight matrix H based on the number of samples in the mini-batches or other data-driven approaches can improve the recognition of well-represented classes.

Table 4.17: Confusion Matrix of Weighted-Loss Approach on the Test Set

		Predicted							
		NE	**HA**	**SA**	**SU**	**FE**	**DI**	**AN**	**CO**
Actual	**NE**	**53.3**	2.8	9.8	8.7	1.7	2.5	10.4	10.9
	HA	4.5	**72.8**	1.1	6.0	0.6	1.7	1.0	12.2
	SA	13.0	1.3	**61.7**	3.6	5.8	4.4	9.2	1.2
	SU	3.4	1.2	1.7	**69.9**	18.9	1.7	2.8	0.5
	FE	1.5	1.5	4.6	13.5	**70.4**	4.2	4.3	0.2
	DI	2.0	2.2	5.8	3.3	6.2	**68.6**	10.6	1.3
	AN	6.2	1.2	5.0	3.2	5.8	11.1	**65.8**	1.9
	CO	16.2	13.1	3.5	3.1	0.5	4.3	5.7	**53.8**

* NE, HA, SA, SU, FE, DI, AN, CO, NO, UN , and NF stand for Neutral, Happy, Sad, Surprise, Fear, Anger, Disgust, Contempt, None, Uncertain, and Non-face categories, respectively.

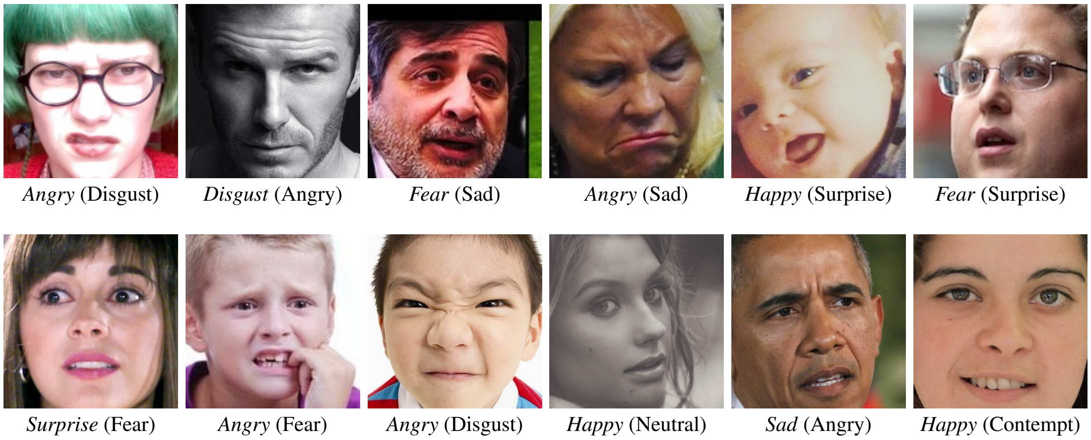
We compared the performance of CNN baseline with a Support Vector Machine (SVM) (Cortes and Vapnik, 1995). To train SVM, the faces in the images were cropped and resized to 256×256 pixels. HOG (Dalal and Triggs, 2005) features were extracted with the cell size of 8. We applied PCA retaining 95% of the variance to reduce the HOG features dimensionality from 36,864 to 6,697 features. We used a linear kernel SVM in Liblinear package (Fan et al., 2008) (which is optimized for large-scale linear classification and regression). Table 4.16 shows the evaluation metrics of SVM. Since Liblinear package does not provide the scores of classifications, the AUC and AUCPR cannot be calculated. Comparing the performance of the SVM with the CNN baselines on AffectNet indicates that CNN models perform better than conventional SVM and HOG features in all metrics.

We also compared the baseline with Microsoft Cognitive emotion API (Microsoft-Oxford, 2015) as an available off-the-shelf expression recognition system. The MS cognitive system had an excellent performance on Neutral and Happy categories with an accuracy of 0.94 and 0.85, respectively. However, it performed poorly on other classes with an accuracy of 0.25, 0.27 and 0.04 in the Fear, Disgust and Contempt categories. Table 4.16 shows the evaluation metrics on the MS cognitive system. Comparing the performance of the MS cognitive with the simple baselines on AffectNet indicates that *AffectNet* is a challenging database and a great resource to further improve the performance of facial expression recognition systems.

Figure 4.15 shows nine samples of randomly selected misclassified images of the weighted-loss approach and their corresponding ground-truth. As the figure shows, it is really difficult to assign some of the emotions to a single category. Some of the faces have partial similarities in facial features to the misclassified images, such as nose wrinkled in disgust, or eyebrows raised in surprise. This emphasizes the fact that classifying facial expressions in the wild is a challenging task and, as mentioned before, even human annotators agreed on only 60.7% of the images.

4.5.3 Dimensional Model (Valence and Arousal) Baseline

Predicting dimensional model in the continuous domain is a real-valued regression problem. We used AlexNet (Krizhevsky et al., 2012) architecture as our deep CNN baseline to predict the value of valence and arousal. Particularly, two separate AlexNets were trained where the last fully-connected layer was replaced with a linear regression layer containing only one neuron. The output of the neuron predicted the value of valence/arousal in



Angry (Disgust) *Disgust* (Angry) *Fear* (Sad) *Angry* (Sad) *Happy* (Surprise) *Fear* (Surprise)

Surprise (Fear) *Angry* (Fear) *Angry* (Disgust) *Happy* (Neutral) *Sad* (Angry) *Happy* (Contempt)

Figure 4.15: Samples of miss-classified images. Their corresponding ground-truth is given in parentheses.

continuous domain [-1,1]. A Euclidean (L2) loss was used to measure the distance between the predicted value ($\hat{y}_n$) and actual value of valence/arousal (y_n) as:

$$E = \frac{1}{2N} \sum_{n=1}^{N} ||\hat{y}_n - y_n||_2^2 \tag{4.23}$$

The faces were cropped and resized to 256×256 pixels. The base learning rate was fixed and set to 0.001 during the training process. We used a momentum of 0.9. Training was continued until a plateau was reached in the Euclidean error of the validation set (approximately 16 epochs with a mini-batch size of 256). Figure 4.16 shows the value of training and validation losses over 16K iterations (about 16 epochs).

We also compared Support Vector Regression (SVR) (Smola and Vapnik, 1997) with our DNN baseline for predicting valence and arousal in *AffectNet*. In our experiments, first, the faces in the images were cropped and resized to 256×256 pixels. Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005) features were extracted with the cell size of 8. Afterward, we applied PCA retaining 95% of the variance of these features to

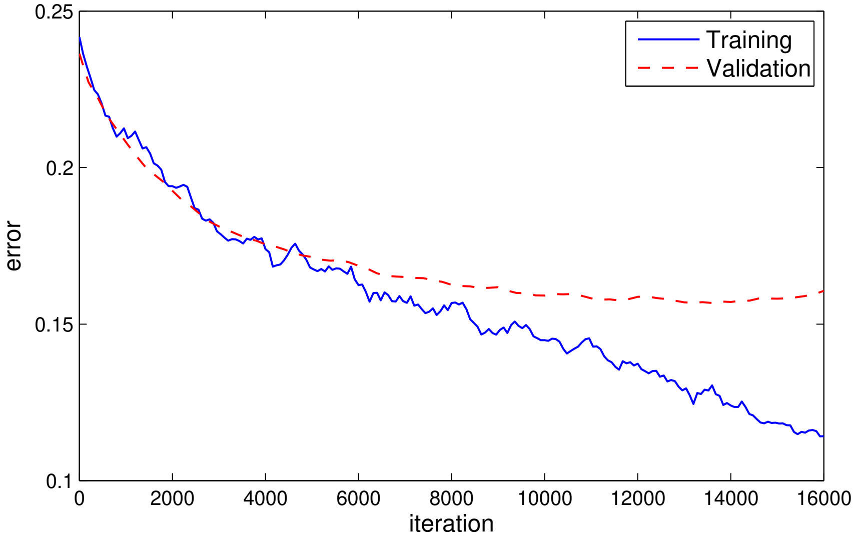


Figure 4.16: Euclidean error of training valence and arousal.

Table 4.18: Baselines' Performances of Predicting Valence and Arousal on Test Set

	CNN (AlexNet)		SVR	
	Valence	Arousal	Valence	Arousal
RMSE	0.394	0.402	0.494	0.400
CORR	0.602	0.539	0.429	0.360
SAGR	0.728	0.670	0.619	0.748
CCC	0.541	0.450	0.340	0.199

* RMSE, CORR, SAGR, and CCC stand for Root Mean Square Error, Correlation, Sign Agreement Metric, and Concordance Correlation Coefficient respectively.

reduce the dimensionality. Two separate SVRs were trained to predict the value of valence and arousal. Liblinear (Fan et al., 2008) package was used to implement SVR baseline.

Table 4.18 shows the performances of the proposed baseline and SVR on the test set. As shown, the CNN baseline can predict the value of valence and arousal better than SVR. This is because the high variety of samples in *AffectNet* allows the CNN to extract more discriminative features than hand-crafted HOG, and therefore it learned a better representation of dimensional affect.

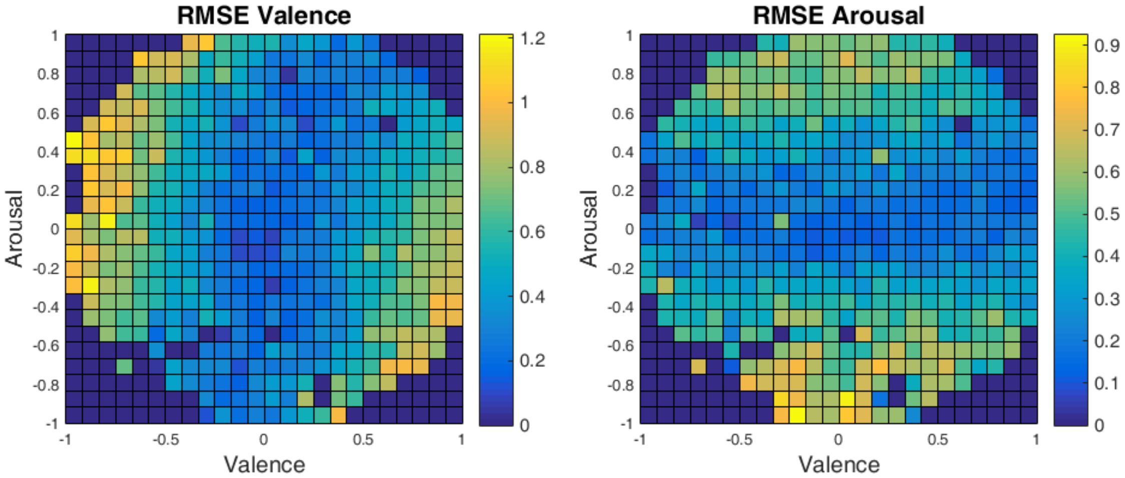


Figure 4.17: RMSE of predicted valence and arousal using AlexNet and Euclidean (L2) loss (Best viewed in color).

The RMSE of CNN baseline (AlexNet) between the predicted valence and arousal and the ground-truth are shown in Fig. 4.17. As illustrated, the CNN baseline has a lower error rate in the center of circumplex. In particular, predicting low-valence mid-arousal and low-arousal mid-valence areas were more challenging. These areas correspond to the expressions of contempt, bored, and sleepy.

4.6 Conclusion

In this chapter, we presented a new deep neural network architecture for automated facial expression recognition. The proposed network consists of two convolutional layers each followed by max pooling and then four Inception layers. The Inception layers increase the depth and width of the network while keeping the computational budget constant. The proposed approach is a single component architecture that takes registered facial images as the input and classifies them into either of the six basic expressions or the neutral.

We evaluated our proposed architecture in both subject-independent and cross-database manners on seven well-known publicly available databases. Our results confirm the superiority of our network compared to several state-of-the-art methods in which engineered features and classifier parameters are usually tuned on a very few databases. Our network is the first work which applies the Inception layer architecture to the FER problem across multiple databases. The clear advantage of the proposed method over conventional CNN methods (i.e. shallower or thinner networks) is gaining increased classification accuracy on both subject independent and cross-database evaluation scenarios while reducing the number of operations required to train the network.

The analysis of human facial behavior is a very complex and challenging problem. The majority of the techniques for automated facial affect analysis are mainly based on machine learning methodologies, and their performance highly depends on the amount and diversity of annotated training samples. Recently, databases of facial expression and affect in the wild received much attention. However, existing databases of facial affect in the wild only cover one model of affect, have a limited number of subjects, or contain few samples of certain emotions.

The Internet is a vast source of facial images, most of which are captured in uncontrolled conditions. These images are often taken in the wild under natural conditions. In this research, we introduced a new publicly available database of a facial **Affect** from the InterNet (called *AffectNet*) by querying different search engines using emotion related tags in six different languages. *AffectNet* contains more than 1M images with faces and extracted landmark points. Twelve human experts manually annotated 440,000 of these images in both the categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face.

The agreement level of human labelers on a subset of *AffectNet* showed that expression recognition and predicting valence and arousal in the wild is a challenging task. The two annotators agreed on 60.7% of the category of facial expressions, and there was a large disagreement on the value of valence and arousal (RMSE=0.34 and 0.36) between the two annotators.

Two simple deep neural network baselines were examined to classify the facial expression images and predict the value of valence and arousal in the continuous domain of dimensional model. Evaluation metrics showed that simple deep neural network baselines trained on *AffectNet* can perform better than conventional machine learning methods and available off-the-shelf expression recognition systems. *AffectNet* is by far the largest database of facial expression, valence and arousal in the wild, enabling further progress in the automatic understanding of facial behavior in both categorical and continuous dimensional space. The interested investigators can study categorical and dimensional models in the same corpus, and possibly co-train them to improve the performance of their affective computing systems. It is highly anticipated that the availability of this database for the research community, along with the recent advances in deep neural networks, can improve the performance of automated affective computing systems in recognizing facial expressions and predicting valence and arousal.

Chapter 5

Affect-Aware Agent

In Chapters 2 and 3, we showed that our proposed rear-projection robotic platform can show natural visual speech and facial expressions, and users preferred the robotic platform over a virtual agent presented on a 2D screen. In Chapter 4, we created a new database of facial expressions and proposed a new automated Facial Expression Recognition (FER) system that is able to recognize users' emotions in facial images in uncontrolled conditions. The goal of this chapter is to integrate emotional intelligence and the proposed automated FER system into spoken dialogs of the developed robotic platform to evaluate whether this integration can add up values to our robot.

The objective of integrating the automated FER system into the robot is to improve social capabilities of the robot and create an expressive and empathic social agent (affect-aware) which is capable of interpreting users' emotional facial expressions and act accordingly. For this purpose, we designed a simple experiment in which the subjects performed a simple in front of the robot, and the robot engages them in conversation based on their perceived facial expressions. We measured the accuracy of the automated FER system on the robot when interacting with different human subjects as well as three social/interaction aspects, namely task engagement, being empathic, and likability of the robot.

The rest of this chapter is organized as follow. Section 5.1 reviews the definition of empathy in the literature and the studies of empathy in social agents. Section 5.2 discusses the experiments and Section 5.3 introduces the questionnaire designed to evaluate the social/interaction aspects of the agent studied in the experiments. Section 5.4 presents our findings on the effect of creating affect-aware robotic agent using automated FER system. Finally, Section 5.5 concludes this chapter.

5.1 Empathy

There are several definitions for Empathy in the literature. These definitions can be divided into three major categories: (1) affective empathy (also called emotional empathy, or primitive empathy), where empathy is an affective response to others' emotional states, (2) cognitive empathy, where empathy is the cognitive understanding of others' emotional states, and (3) combination of both affective and cognitive components (Omdahl, 1995; Staub, 1987). Researchers such as Davis (1994), Hoffman (2001), and Preston and De Waal (2002) have attempted to unify different perspectives of empathy by adopting a multidimensional approach. Davis (1994) defines empathy as:

> *"A set of constructs having to do with the responses of one individual to the experiences of another"*.

Similarly, Hoffman (2001) defines empathy as a psychological process that makes a person have:

> *"Feelings that are more congruent with another's situation than with his own situation"*.

Preston and De Waal (2002) defines empathy as a “Perception-Action Model” (PAM) that considers several phenomena (and processes) such as emotional contagion and sympathy. They defined empathy as the capacity to (a) be affected by and share the emotional state of another, (b) assess the reasons of emotional state, and (c) identify and adopt other perspectives. In this work, we adopt this definition of empathy. Along with this definition, six elements are involved in an empathic situation (Paiva et al., 2017):

1. **Observer (empathizer)**: a person/agent who responds emotionally to the affective state of another one.
2. **Target**: a person/agent who expresses an emotional state (or is in an emotional situation perceived by the observer).
3. **Event**: an event that happens and is witnessed by the observer.
4. **Emotion**: an emotion to which the observer responds.
5. **Context**: a context/situation where the event happens.
6. **Mediating factors**: factors such as the relation between the observer and the target, the mood of the observer, the presence or absence of another person/agent, past situations, etc.

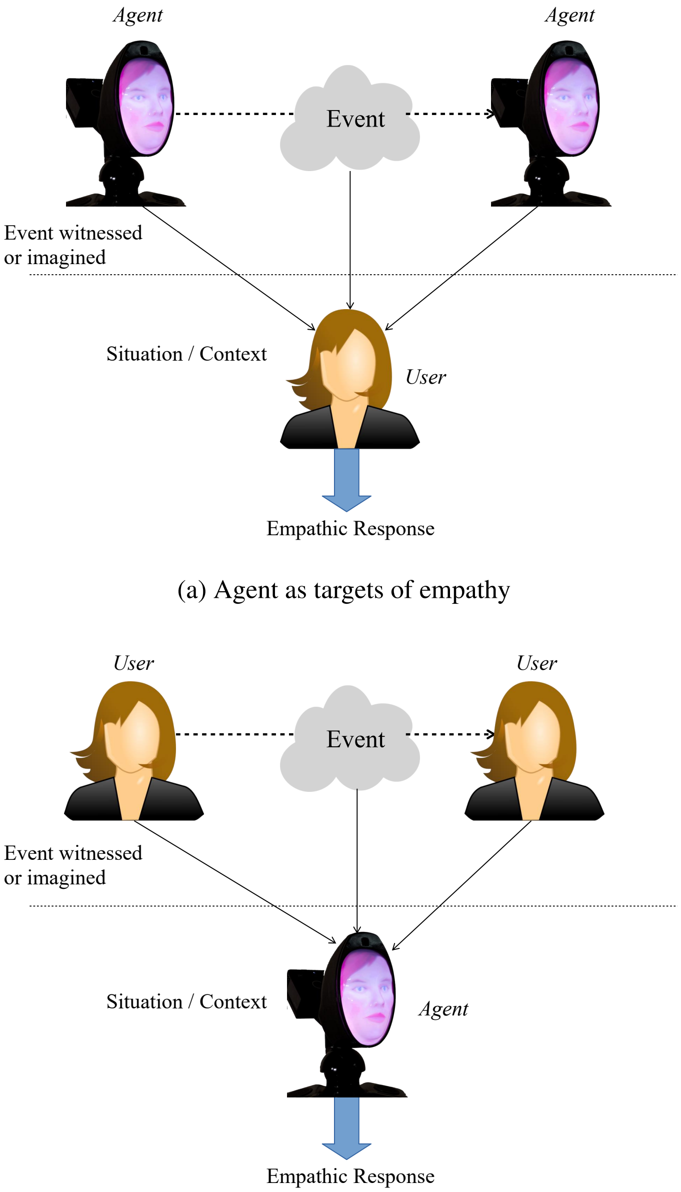
5.1.1 Empathy in Social Assistive Robots

Studies on empathy in social agents can be divided into two approaches (Paiva et al., 2017). The first approach is to consider the human as the observer (empathizer) and the agent as the target that triggers empathy in the human partner (see Figure 5.1a). In this case, the agent does not necessarily show empathic behavior, but it is designed to evoke empathy in the human observer. These agents are used for helping children to deal with bullying

(Paiva et al., 2004), training young doctors on interview patients with depression (Marsella et al., 2000), and intercultural training with culturally configurable agents (Mascarenhas et al., 2013). The second approach—and the focus of this work—is to build agents with empathic behavior, i.e., agents as observers that empathize with human partners as targets of empathy (see Figure 5.1b).

Developing agents with empathic behavior has received considerable attention, especially in the virtual agent community (Paiva et al., 2017). Studies have shown that empathic virtual agents are perceived as more likable, trustworthy and caring (Brave et al., 2005), reduce stress (Prendinger and Ishizuka, 2005), and they can build and sustain long-term socio-emotional relationships with human partners (Bickmore and Picard, 2005). In the field of social robotics, however, researchers have only more recently started to assess the effects of empathy in human-robot interaction (compared to the research on the virtual agents). This is mainly due to the advances and improvement in automatic affect recognition in different modalities (e.g., facial expression recognition, speech tone sentiment analysis, speech-to-text sentiment analysis etc.), which enables robots to interact with users in a less controlled setting (i.e., user facial pose, environment illumination/noise etc.).

Several works in the area of social robotics have studied the effect of evoking empathy in users (Hayes et al., 2014; Kwak et al., 2013; Rosenthal-von der Pütten et al., 2013; Seo et al., 2015). Majority of these works concluded that users have more empathy with a physical robot than a virtual agent. For example, Kwak et al. (2013) studied the effect of evoking empathy in users for different levels of agency (i.e., a mediated robot which delivers the emotional state of a remote user or a simulated robot which expresses its own emotion) and the physical embodiment (i.e., physically embodied or physically disembodied). Inspired by Milgram (1963) experiment, participants were allowed to punish the robot with an electric shock when the robot responded incorrectly, and the robot showed some type of "bruises" through light-emitting diodes to express negative emotional states. The



(a) Agent as targets of empathy

(b) Empathic agents as observer (focus of this study)

Figure 5.1: Two different perspectives of studying empathy in social agents (Paiva et al., 2017).

results indicated that participants empathized more with the mediated robot than with the simulated robot. Also, the participants empathized more with a physically embodied robot than with a physically disembodied one.

In another interesting work, Seo et al. (2015) evaluated how people empathize with a physical or simulated virtual agent of the robot when something bad happens to it. Participants played a collaborative game with NAO robot. At some point, the robot which is controlled by WoZ started expressing "problems" due to a virus, which made it work badly and eventually, the fault cleared the agent's memory. Their results suggested that people empathize more with a physical robot than a simulated one. A detailed comparison of different studies that used robot/virtual agents to evoke users' empathy can be found in a recent survey (Paiva et al., 2017).

There are few works that used social robots as observers that empathize with human partners as targets of empathy (see Figure 5.1b). Riek et al. (2010) investigated how imitation by a robot can affect people's perceptions of their conversation with it. The robot operated in one of three ways: full head gesture mimicking, partial head gesture mimicking (nodding), and non-mimicking. Participants engaged in two conversational tasks with the robot, one non-emotional (i.e., describe the route they took to the laboratory that day), and the other one emotionally salient (i.e., tell their first memories of Cambridge—people they met, things they saw, foods they ate, etc.). After the experiment, participants rated a modified version of the Interactant Satisfaction Survey (Kang et al., 2008) with fifteen item that measured the social attraction toward and emotional credibility of conversation partners. The results indicated that the participants in the full head gesture condition rated their interaction the most positively, followed by the partial and non-mimicking conditions.

Cramer et al. (2010) studied how empathy affects people's attitudes towards robots. In a between-subjects design, two groups of subjects participated in an online survey experiment and watched a four-minute video of an actor playing a cooperative game with an iCat robot (Van-Breemen, 2004). The robot expressed empathic behavior towards the actor in three conditions: emphatically accurate, neutral, and inaccurate (i.e., incongruent behavior to the situation). The study showed that subjects' trust decreased when robot's empathic responses were incongruent with the affective state of the subjects. Contrarily, subjects who observed the robot displaying accurate empathic behaviors felt a higher relationship with the robot.

Leite et al. (2014) studied empathic model for social robots aiming to interact with children for extended periods of time. Sixteen subjects from the 3rd grade played a total of five chess exercises with the iCat robot over five consecutive weeks. After every child's movement on the chessboard, the robot provided empathic feedback on that move by conveying facial expressions influenced by the child's affective state (positive, negative, and neutral) and the state of the game. In addition, if the child's affective state is negative and below a certain threshold, the robot also displayed social supportive behaviors. After playing with the robot, in the first and last weeks of interaction children filled in a questionnaire. The result indicated that the ratings of social presence, engagement and self-validation remained similar after five weeks, contrasting with a similar study (Leite et al., 2009) where the robot was not endowed with the empathic model.

5.2 Methodology

In this work, we aim to integrate our developed FER module to our robotic platform and study the role of understanding user's facial expression in three social/interaction aspects (task engagement, being empathic, and likability of the robot). For this purpose, we studied three robotic agent conditions:

1. **Non-Empathic (NE):** In this condition, the robot does not recognize user's affect (or assume the user's affect is neutral) and only performs the task.

2. **Automated Recognition-Empathic (AR-E):** In this condition, the developed automated FER system is used to recognize user's affect. If the recognized affect is not neutral, the robot empathizes with the user via a set of predefined conversation.

3. **Human Recognition-Empathic (HR-E):** In this condition, instead of automated FER system, a human observer recognizes user's affect and prompts the robot. If the recognized affect is not neutral, the robot empathizes with the user via the same set of conversation used in the AR-E condition. The purpose of performing the experiments with this condition is to evaluate the response of the users to an optimal case (optimal perception of facial expression) in our research setting.

In order to evoke emotion in subjects, two stimuli each containing four video clips (approximately 30 seconds long) were created. Each clip was intended to elicit a certain spontaneous emotion (i.e., happy, surprise, sad, disgust) taken mostly from YouTube. The clips contained segments of videos such as laughing baby and funny scenes to evoke happiness, survival of car crashes to evoke surprise, dogs crying over dead owner/friends to evoke sadness, and eating giant larva and dead animal to evoke the emotion of disgust. We did not include fear and anger (two of six basic facial expressions) in the stimuli and experiment, as scaring subject may emotionally disturb them and it is difficult to evoke anger with a short video. Further information about the video clip is provided in the Appendix C.1.

Subjects were not fully aware of the experiment's intention (i.e., empathizing with subjects), and the task of the experiment was described to them as "You watch some videos and the robot will ask you to describe the video in one word". All participants in the study were IRB consented prior to enrolling in the experiment. Participation was completely voluntary and the subjects were told that they may feel sad or disgusted while watching some of the videos and they can leave the study at any time. While watching the videos, subjects sat on a comfortable chair positioned in front of a 19" LCD display and the robot at a distance of

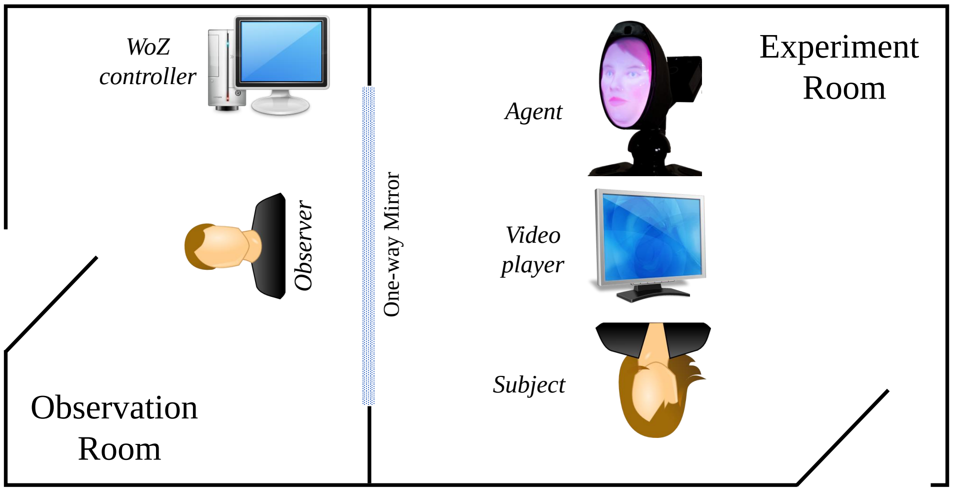


Figure 5.2: Evaluation of empathic agent room setup

60 cm in all three conditions. They were alone in the experiment room while an observer was watching them through a one-way mirror (See Figure 5.2). The observer recorded the facial expression of the subjects during watching the videos. The facial behaviors of subjects were also video recorded for further analysis.

In the beginning of the experiment, Ryan introduced herself and explained the task as: “Hello my friend. My name is Ryan. We are going to have a short experiment together. In this experiment, you are going to watch some videos and I appreciate it if you describe the video in one word to me. Are you ready?”. The robot waits until the user is ready, and then displays one of the stimuli randomly with the fix sequence of happy, surprise, sad and disgust in all three conditions (NE, AR-E, and HR-E). We chose this sequence, i.e., showing clips intended to elicit positive emotion first, hence it might be difficult (or even impossible) to evoke positive emotions with a short video after emotionally disturbing subjects with negative emotions. Figure 5.3 shows the conversation graph of the experiment.

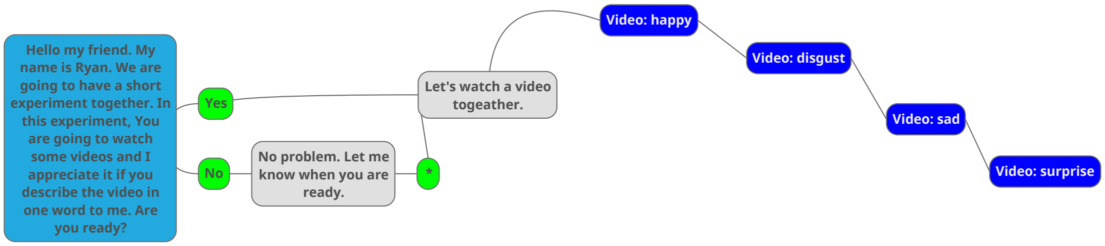

Hello my friend. My name is Ryan. We are going to have a short experiment together. In this experiment, You are going to watch some videos and I appreciate it if you describe the video in one word to me. Are you ready?
Yes
No
No problem. Let me know when you are ready.
*
Let's watch a video togeather.
Video: happy
Video: disgust
Video: sad
Video: surprise

Figure 5.3: Conversation graph of the experiment

5.2.1 Automated FER System

In order to enhance the accuracy of the proposed deep neural network baseline model for FER (Section 4.5) and since we only study four facial expressions in this work, we trained a 50-layers Residual Network (ResNet) (He et al., 2016) on five classes of neutral, happy, surprise, sad, and disgust of affectNet database (Section 4.4). The ResNet architecture is a state-of-the-art CNN with added shortcut connections, i.e., a linear transform of each layer's input to the layer's output. Figure 5.4 shows a building block of residual learning in ResNet. Adding the shortcut connection eases the training of deeper networks (more than 100 layers) and avoid degradation problem (the phenomenon that accuracy gets saturated and then degrades rapidly (He et al., 2016)). The residual connection has yielded state-of-the-art performance in several computer vision applications such as visual object detection (He et al., 2016), semantic image segmentation (Chen et al., 2016), audio classification (Hershey et al., 2017), and facial expression recognition (Hasani and Mahoor, 2017).

During the experiments, subjects' faces were captured by a webcam installed on the video player monitor. The OpenCV face recognition library was used to detect faces in the images, and 66 landmark points were found using a face alignment algorithm via regression local binary features (Ren et al., 2014; Yu, 2016). We used these points to register faces to an average face using an affine transformation. Once the faces have been registered, the

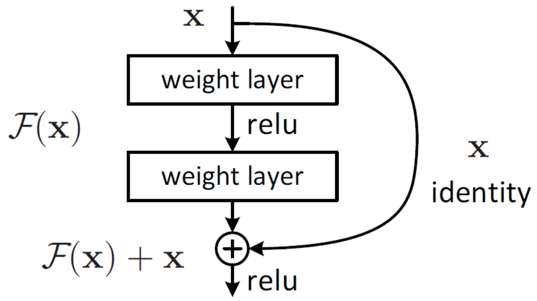


Figure 5.4: Building block of Residual learning in ResNet (He et al., 2016)

face regions were cropped, resized to 48×48 pixels, and fed into the trained network for classification.

We used a K40 GPU for training the network, and an Intel Core i7 CPU during the inference. The face detection, registration, and expression classification take ∼20ms, enabling us to process five frames per second. A majority voting is used to determine user's facial expression during watching the video. As videos trigger emotions in few scenes and users had neutral faces in the rest of time of watching the videos, the frames with emotions detected as neutral faces were discarded by the probability of 0.5 in a majority voting scheme.

5.2.2 Empathic Conversations

According to Preston and De Waal (2002) empathy reaction can be a function of three factors:

1. Be affected by and share the emotional state of another.

2. Assess the reasons of emotional state.

3. Identify and adopt other perspectives.

Taking into account these elements and the previously given definition of empathy, we propose the following features that need to be embodied in our empathic robot:

- The robot should be capable of recognizing, understanding, and interpreting the user's emotional state (facial expression in this experiment).
- The robot should be capable of expressing its emotion using both verbal and non-verbal cues.
- The robot should be capable of perspective taking, being supportive, and have self-correction to adopt other perspectives.

The robot recognizes user's facial expression during watching the videos. Based on the affective state of the user, the robot appraises the situation and generates empathic responses, e.g., "congruent facial expressions" in tune with the user's affective state, "perspective-taking", "being supportive", and "self-correction".

A set of predefined empathic responses based on the perceived affect state and conversation with users were carefully designed. Figures 5.5, 5.6, 5.7, and 5.8 show empathic conversation map after showing videos intended to elicit happy, surprise, sad, and disgust emotions, respectively. For instance, as shown in Figure 5.7, if Ryan recognizes sadness, she shows sad face [congruent facial expressions] and says *"It looks like the video made you sad. Am I right?"*. If the user confirms that he/she was sad, the robot keeps the sad face and say *"Me too. It was heartbreaking. I am sorry you had to watch it."* [perspective-taking], *"Do you want to talk about it?"*. Based on user's response, the robot will say *"I understand. It was obvious from your face."* [perspective- taking], *"I wish I could hug you"* [being supportive]. If the user did not have a negative affect (e.g., user had a neutral face) and the robot recognized it incorrectly, the robot stops showing sad face and says *"Oh. Seems like I misinterpreted your face"* [self-correction], *"You are focused on the task. Good Job"* [being supportive].

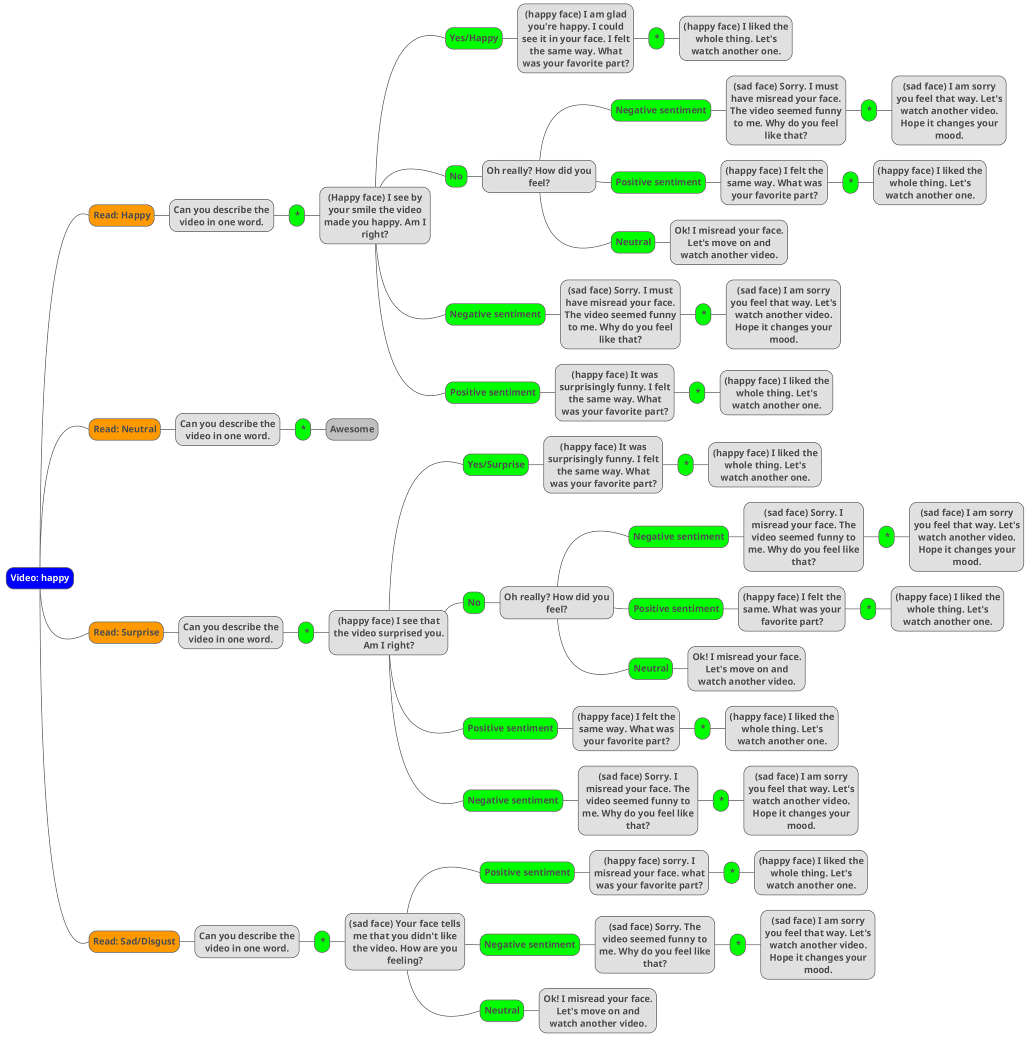

Video: happy
Read: Happy
Can you describe the video in one word.
*
(Happy face) I see by your smile the video made you happy. Am I right?
Yes/Happy
(happy face) I am glad you're happy. I could see it in your face. I felt the same way. What was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
No
Oh really? How did you feel?
Negative sentiment
(sad face) Sorry. I must have misread your face. The video seemed funny to me. Why do you feel like that?
*
(sad face) I am sorry you feel that way. Let's watch another video. Hope it changes your mood.
Positive sentiment
(happy face) I felt the same way. What was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
Neutral
Ok! I misread your face. Let's move on and watch another video.
Negative sentiment
(sad face) Sorry. I must have misread your face. The video seemed funny to me. Why do you feel like that?
*
(sad face) I am sorry you feel that way. Let's watch another video. Hope it changes your mood.
Positive sentiment
(happy face) It was surprisingly funny. I felt the same way. What was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
Read: Neutral
Can you describe the video in one word.
*
Awesome
Read: Surprise
Can you describe the video in one word.
*
(happy face) I see that the video surprised you. Am I right?
Yes/Surprise
(happy face) It was surprisingly funny. I felt the same way. What was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
No
Oh really? How did you feel?
Negative sentiment
(sad face) Sorry. I misread your face. The video seemed funny to me. Why do you feel like that?
*
(sad face) I am sorry you feel that way. Let's watch another video. Hope it changes your mood.
Positive sentiment
(happy face) I felt the same. What was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
Neutral
Ok! I misread your face. Let's move on and watch another video.
Positive sentiment
(happy face) I felt the same way. What was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
Negative sentiment
(sad face) Sorry. I misread your face. The video seemed funny to me. Why do you feel like that?
*
(sad face) I am sorry you feel that way. Let's watch another video. Hope it changes your mood.
Read: Sad/Disgust
Can you describe the video in one word.
*
(sad face) Your face tells me that you didn't like the video. How are you feeling?
Positive sentiment
(happy face) sorry. I misread your face. what was your favorite part?
*
(happy face) I liked the whole thing. Let's watch another one.
Negative sentiment
(sad face) Sorry. The video seemed funny to me. Why do you feel like that?
*
(sad face) I am sorry you feel that way. Let's watch another video. Hope it changes your mood.
Neutral
Ok! I misread your face. Let's move on and watch another video.

Figure 5.5: Empathic conversation map after showing a video intended to elicit a happy emotion

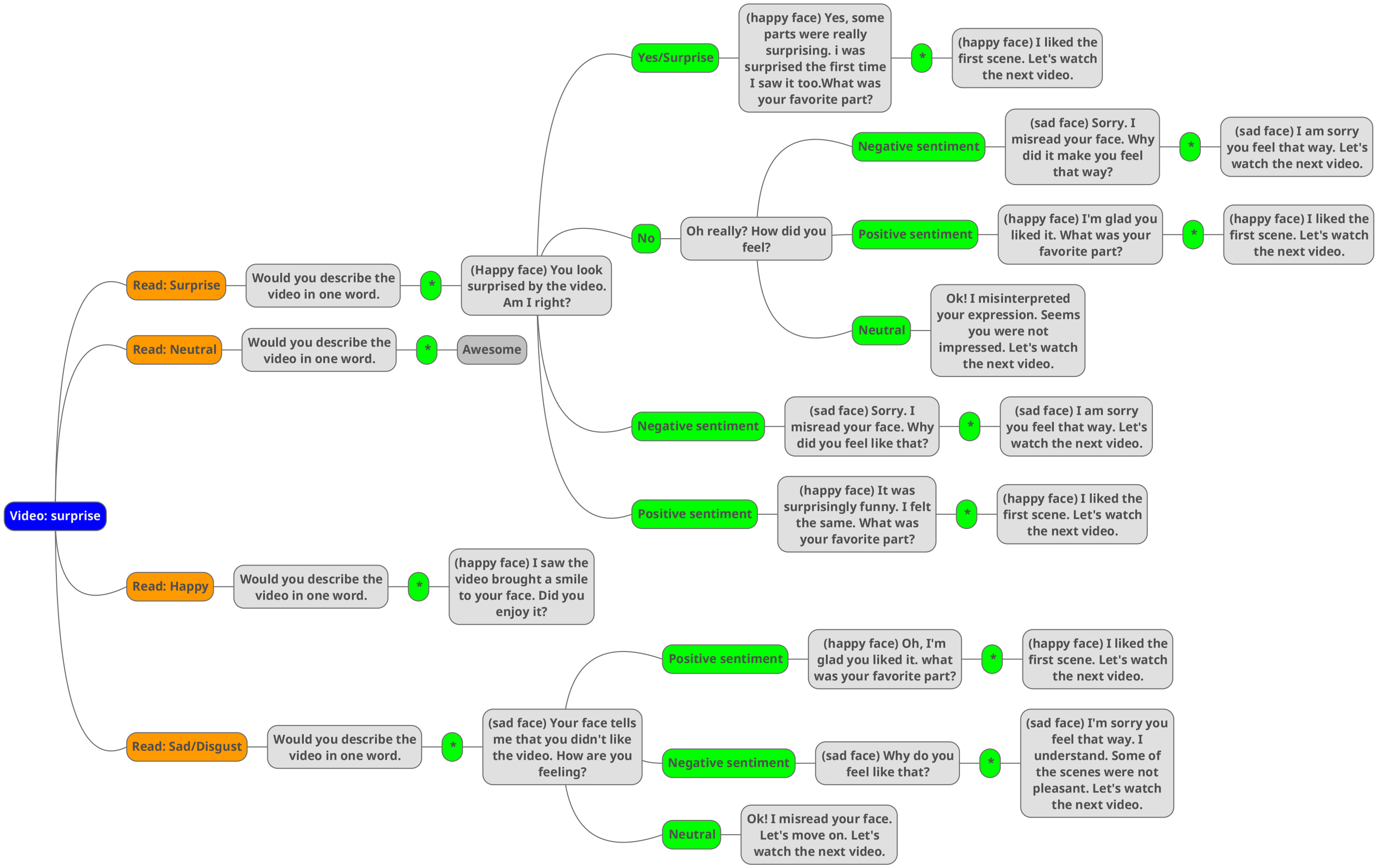

Video: surprise
Read: Surprise
Would you describe the video in one word.
*
(Happy face) You look surprised by the video. Am I right?
Yes/Surprise
(happy face) Yes, some parts were really surprising. i was surprised the first time I saw it too.What was your favorite part?
*
(happy face) I liked the first scene. Let's watch the next video.
No
Oh really? How did you feel?
Negative sentiment
(sad face) Sorry. I misread your face. Why did it make you feel that way?
*
(sad face) I am sorry you feel that way. Let's watch the next video.
Positive sentiment
(happy face) I'm glad you liked it. What was your favorite part?
*
(happy face) I liked the first scene. Let's watch the next video.
Neutral
Ok! I misinterpreted your expression. Seems you were not impressed. Let's watch the next video.
Negative sentiment
(sad face) Sorry. I misread your face. Why did you feel like that?
*
(sad face) I am sorry you feel that way. Let's watch the next video.
Positive sentiment
(happy face) It was surprisingly funny. I felt the same. What was your favorite part?
*
(happy face) I liked the first scene. Let's watch the next video.
Read: Neutral
Would you describe the video in one word.
*
Awesome
Read: Happy
Would you describe the video in one word.
*
(happy face) I saw the video brought a smile to your face. Did you enjoy it?
Read: Sad/Disgust
Would you describe the video in one word.
*
(sad face) Your face tells me that you didn't like the video. How are you feeling?
Positive sentiment
(happy face) Oh, I'm glad you liked it. what was your favorite part?
*
(happy face) I liked the first scene. Let's watch the next video.
Negative sentiment
(sad face) Why do you feel like that?
*
(sad face) I'm sorry you feel that way. I understand. Some of the scenes were not pleasant. Let's watch the next video.
Neutral
Ok! I misread your face. Let's move on. Let's watch the next video.

Figure 5.6: Empathic conversation map after showing a video intended to elicit a surprise emotion

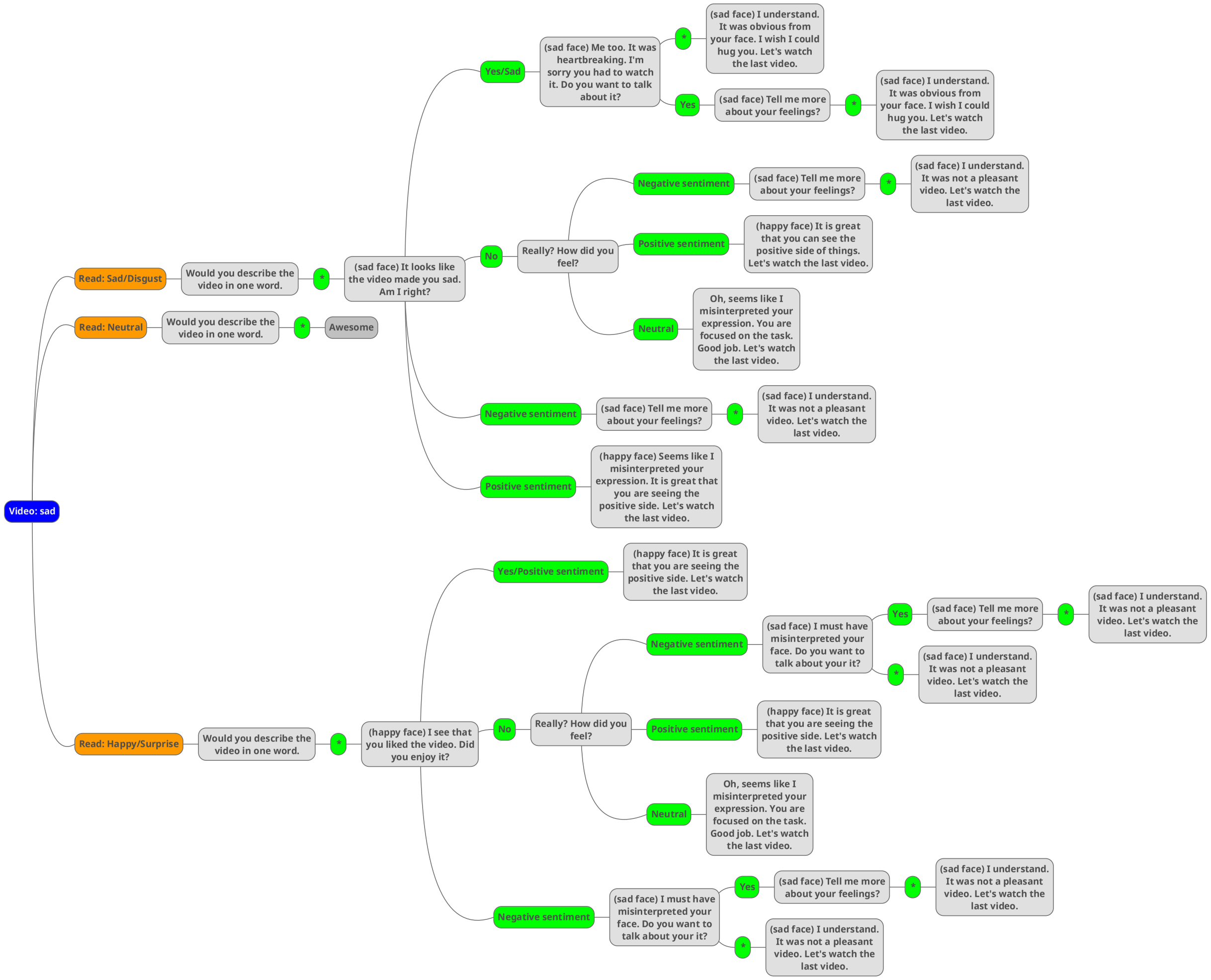

Video: sad
Read: Sad/Disgust
Would you describe the video in one word.
*
(sad face) It looks like the video made you sad. Am I right?
Yes/Sad
(sad face) Me too. It was heartbreaking. I'm sorry you had to watch it. Do you want to talk about it?
*
(sad face) I understand. It was obvious from your face. I wish I could hug you. Let's watch the last video.
Yes
(sad face) Tell me more about your feelings?
*
(sad face) I understand. It was obvious from your face. I wish I could hug you. Let's watch the last video.
No
Really? How did you feel?
Negative sentiment
(sad face) Tell me more about your feelings?
*
(sad face) I understand. It was not a pleasant video. Let's watch the last video.
Positive sentiment
(happy face) It is great that you can see the positive side of things. Let's watch the last video.
Neutral
Oh, seems like I misinterpreted your expression. You are focused on the task. Good job. Let's watch the last video.
Negative sentiment
(sad face) Tell me more about your feelings?
*
(sad face) I understand. It was not a pleasant video. Let's watch the last video.
Positive sentiment
(happy face) Seems like I misinterpreted your expression. It is great that you are seeing the positive side. Let's watch the last video.
Read: Neutral
Would you describe the video in one word.
*
Awesome
Read: Happy/Surprise
Would you describe the video in one word.
*
(happy face) I see that you liked the video. Did you enjoy it?
Yes/Positive sentiment
(happy face) It is great that you are seeing the positive side. Let's watch the last video.
No
Really? How did you feel?
Negative sentiment
(sad face) I must have misinterpreted your face. Do you want to talk about your it?
Yes
(sad face) Tell me more about your feelings?
*
(sad face) I understand. It was not a pleasant video. Let's watch the last video.
*
(sad face) I understand. It was not a pleasant video. Let's watch the last video.
Positive sentiment
(happy face) It is great that you are seeing the positive side. Let's watch the last video.
Neutral
Oh, seems like I misinterpreted your expression. You are focused on the task. Good job. Let's watch the last video.
Negative sentiment
(sad face) I must have misinterpreted your face. Do you want to talk about your it?
Yes
(sad face) Tell me more about your feelings?
*
(sad face) I understand. It was not a pleasant video. Let's watch the last video.
*
(sad face) I understand. It was not a pleasant video. Let's watch the last video.

Figure 5.7: Empathic conversation map after showing a video intended to elicit a sad emotion

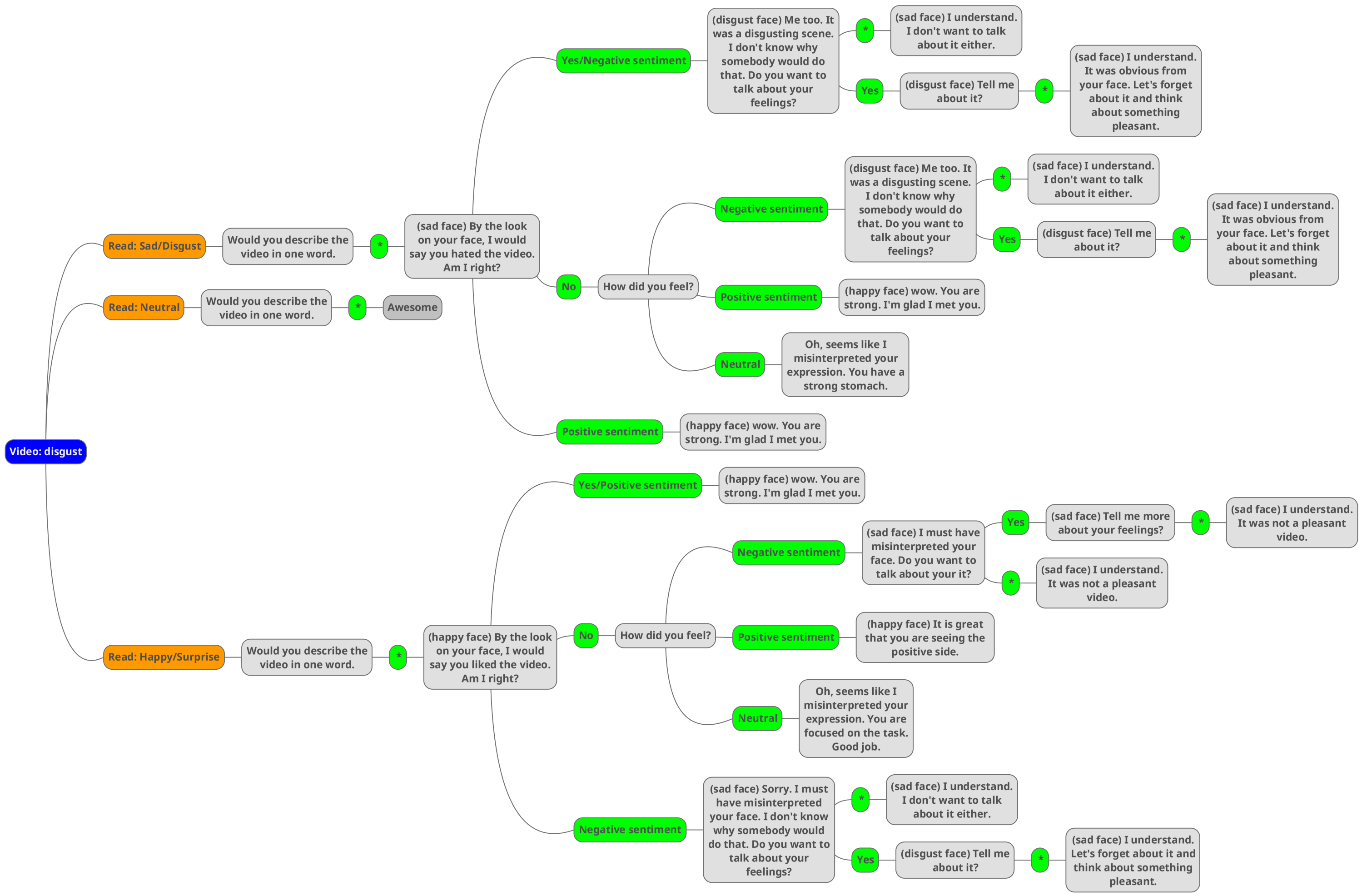

Video: disgust
Read: Sad/Disgust
Would you describe the video in one word.
*
(sad face) By the look on your face, I would say you hated the video. Am I right?
Yes/Negative sentiment
(disgust face) Me too. It was a disgusting scene. I don't know why somebody would do that. Do you want to talk about your feelings?
*
(sad face) I understand. I don't want to talk about it either.
Yes
(disgust face) Tell me about it?
*
(sad face) I understand. It was obvious from your face. Let's forget about it and think about something pleasant.
No
How did you feel?
Negative sentiment
(disgust face) Me too. It was a disgusting scene. I don't know why somebody would do that. Do you want to talk about your feelings?
*
(sad face) I understand. I don't want to talk about it either.
Yes
(disgust face) Tell me about it?
*
(sad face) I understand. It was obvious from your face. Let's forget about it and think about something pleasant.
Positive sentiment
(happy face) wow. You are strong. I'm glad I met you.
Neutral
Oh, seems like I misinterpreted your expression. You have a strong stomach.
Positive sentiment
(happy face) wow. You are strong. I'm glad I met you.
Read: Neutral
Would you describe the video in one word.
*
Awesome
Read: Happy/Surprise
Would you describe the video in one word.
*
(happy face) By the look on your face, I would say you liked the video. Am I right?
Yes/Positive sentiment
(happy face) wow. You are strong. I'm glad I met you.
No
How did you feel?
Negative sentiment
(sad face) I must have misinterpreted your face. Do you want to talk about your it?
Yes
(sad face) Tell me more about your feelings?
*
(sad face) I understand. It was not a pleasant video.
*
(sad face) I understand. It was not a pleasant video.
Positive sentiment
(happy face) It is great that you are seeing the positive side.
Neutral
Oh, seems like I misinterpreted your expression. You are focused on the task. Good job.
Negative sentiment
(sad face) Sorry. I must have misinterpreted your face. I don't know why somebody would do that. Do you want to talk about your feelings?
*
(sad face) I understand. I don't want to talk about it either.
Yes
(disgust face) Tell me about it?
*
(sad face) I understand. Let's forget about it and think about something pleasant.

Figure 5.8: Empathic conversation map after showing a video intended to elicit a disgust emotion

5.3 Social/Interaction Aspects Measurements

We evaluated the impact of the affect recognition system in three social/interaction aspects, i.e., task engagement, empathy, and likability of the robot. For this purpose, we designed a 23-items questionnaire on a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". The questionnaire is shown in Appendix C (Table C.1). The questions are designed as follows:

Task engagement: Questions 1-8 are designed to assess the engagement and enjoyment of the task. Examples of these questions are: *"The task of describing the videos was easy for me."*, *"I enjoyed watching the videos."*, *"The experiment was exciting."*, *"I felt mentally immersed in the experiment."*, etc.

Empathy of the robot: Questions 9-19 measure the empathy of the robot, borrowed from EMOTE project questionnaire. The EMOTE project was a three year FP7 research project which ended in March 2016 (Project, 2013). The EMOTE project questionnaire is inspired by some of the dimensions of Interpersonal Reactivity Index (IRI) (Davis, 1983) where it is also claimed to be a proper resource for measuring empathy in socially assistive robots (Tapus and Mataric, 2007). In the EMOTE questionnaire, instead of asking about the user's perceptions of their own empathic capacities, the items aimed at appraising the robot's empathy. The EMOTE questionnaire has been used in short and long-term studies in the EMOTE project, both for individual interactions (one person and one robot) and multiuser interactions (Paiva et al., 2017).

The EMOTE questionnaire has 14 questions. We excluded three questions in our questionnaire as they did not fit to our task, e.g., "<the robot> tries to look at all sides of an issue before he makes a decision". Examples of questions 9 to 19 in our questionnaire are: *"Ryan can have tender and concerned feelings for people less fortunate than her-*

self.”, *‘Ryan sometimes tried to understand me better by imagining how things look from my perspective.”*, *“Ryan tries to imagine how she would feel if she was in my place.”*, etc.

Likability of the robot: Questions 20-23 estimate the likability of the robot, borrowed from Reysen Likability Scale (Reysen, 2005). Reysen Likability Scale is an 11-items questionnaire on a 7-point Likert scale ranging from “Very Strongly Disagree” to “Very Strongly Agree” which is intended to measure the perceived likability of a target individual. Reysen Likability Scale has been extensively used in robotic literature to evaluate likability of a robot in different scenarios (Li et al., 2010; Rau et al., 2009; Riek and Robinson, 2011). Due to the limitation in the number of questions and in order to be consistent with other questions, we only selected four questions from Reysen Likability Scale and changed the scale to 5-point Likert scale ranging from “Strongly Disagree” to “Strongly Agree”. The selected questions are: *“The robot was friendly.”*, *‘The robot was likable.”*, *“The robot was warm.”*, and *“The robot was approachable.”*.

5.4 Results

We evaluated our proposed empathy model with 16 typical adults, 9 female and 7 males, with age range 18-35 years (Mean= 25.6, SD=5.5) and a variety of ethnicities (12 Caucasian, two Asian, one Hispanic, and one Black). Eight subjects participated in AR-E and NE (i.e., control) conditions, and the other eight subjects performed the experiment with HR-E and NE conditions. The order of conditions was randomly assigned to each subject.

Subject’s facial expressions were recorded by a human observer in the observation room as well as automated FER system in all conditions. In NE condition, robot assumed neutral facial expressions. In HR-E condition, the robot picked human recognized user’s facial expression. In the AR-E condition, the robot acted based on automated FER system. The accuracy of automated facial expression with respect to recorded emotions by the human

Table 5.1: Confusion Matrix of Automated Facial Expression Recognition Accuracy (%)

		Automated FER system				
		Neutral	Happy	Surprise	Sad	Disgust
Human Recognized	Neutral	**28.0**	20.0	36.0	14.0	2.0
	Happy	2.9	**82.9**	14.3	0.0	0.0
	Surprise	7.1	28.6	**50.0**	7.1	7.1
	Sad	23.5	29.4	5.9	**35.3**	5.9
	Disgust	10.0	30.0	**35.0**	20.0	5.0

Table 5.2: Subjects' facial expression recorded by the human observer during watching videos (%)

		Subject Facial Expression				
		Neutral	Happy	Surprise	Sad	Disgust
Video	Happy	11.8	**88.2**	0.0	0.0	0.0
	Surprise	**58.8**	0.0	41.12	0.0	0.0
	Sad	**50.0**	0.0	0.0	**50.0**	0.0
	Disgust	26.5	14.7	0.0	0.0	**58.8**

observer was 49%. Table 5.1 shows the confusion matrix between recorded emotions by the human observer and automated FER system.

Table 5.2 illustrates facial expressions expressed by subjects (recorded by the human observer) for each video clip intended to elicit different emotions. As shown, the videos intended to elicit happiness were most successful in evoking the intended emotion (88.2%), while the videos intended to elicit surprise were least successful (41.12%) and the majority of the subjects remained neutral during watching these videos.

The subjects filled out the questionnaire after each condition. Figure 5.9 shows the average subjects' rating and standard error of each condition. We performed a 3 (Agent Conditions; NE, AR-E and HR-E) × 3 (Aspects; robot's likability, empathy and task's engagement) ANOVA with social/interaction aspects as the within-subject factor and agent condition as the between-subjects factor. The dependent variable was average subject's rating for each social/interaction aspect. The test showed a significant main effect of agent condition [$F(2,29)$ = 17.49, p <.0001, h_p^2 = .547] and a significant main effect of social/interaction

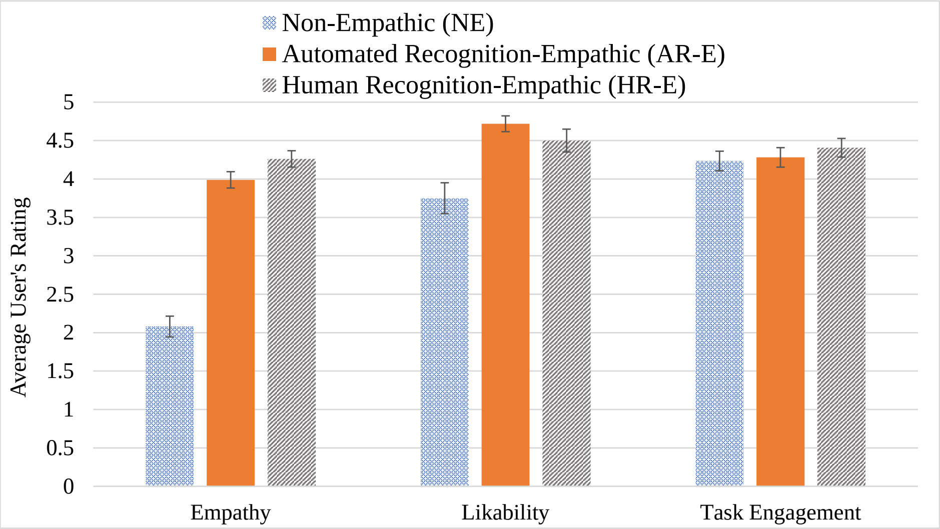


Figure 5.9: The average user's rating and standard error of different empathic agents.

aspect [$F(2, 58)$ = 14.29, p <.0001, h_p^2 = .330]. The interaction between agent condition and social/interaction aspect was also significant [$F(4, 58)$ = 7.84, p <.0001, h_p^2 = .351].

To inspect whether this significant difference exists among all conditions, we performed pairwise two-tailed t-test comparisons between different conditions for different aspects. Table 5.3 shows pairwise p-value and Cohen's d effect-size between agent conditions. As mentioned in Chapter 3, Cohen's d is an effect size used to indicate the standardized difference between two groups defined in Equation (3.2). Generally, the effect size is considered small if $d > 0.2$, medium if $d > 0.5$ and large if $d > 0.8$ (Cohen, 1977).

As indicated in Tables 5.3:

- Subjects rated empathy and likability of both AR-E and HR-E agents more than NE. This shows that the predefined set of conversations and recognizing user's facial expressions improved the user ratings of empathy and likability.

- There was no significant difference between subjects' ratings of empathy and likability in AR-E and HR-E. This shows that the subject found the agent endowed with

Table 5.3: Pairwise comparison (T-test p-value) and Cohen's d effect size of users' rating for different empathic agents. Significant pairs are shown in bold.

	NE vs AR-E		NE vs HR-E		AR-E vs HR-E	
	p	d	p	d	p	d
Empathy	**<.001**	**1.143**	**<.001**	**1.908**	0.072	0.272
Likability	**<.001**	**0.915**	**<.001**	**0.75**	0.231	0.302
Task Engagement	0.370	0.062	0.116	0.172	0.477	0.125

* NE, AR-E, and HR-E stand for Non-Empathic, Automated Recognition-Empathic, and Human Recognition-Empathic robots, respectively.

the automatic FER as good as human recognized expressions, despite the lower accuracy of automated FER. We believe that "self-correction" played an important role in compensating poor accuracy of automated FER system.

- There was no significant difference between subjects' ratings of the task in all conditions. In other word, subjects rated the task engagement similar in all conditions, regardless of being empathized or not.

5.5 Conclusion

We extended and enriched the capabilities of our proposed robotic platform beyond spoken dialogs to create a system that can measure and infer users' affect and cognition. We integrated our proposed automated FER system into spoken dialogs of our robotic platform and evaluated whether this integration can improve three social/interaction aspects (task engagement, being empathic, and likability of the robot). For this purpose, we designed a simple experiment in which the subjects watched some videos to evoke their emotions and the robot asked them to describe each video in one word. During watching the videos, the robot recognized subjects' facial expressions and engaged them in conversation based on the perceived facial expressions.

We studied three conditions as:

- Non-Empathic robot (NE), in which the robot did not recognize user's affect and only performed the task (control condition).
- Automated Recognition-Empathic (AR-E), in which the developed automated FER system was used to recognize subject's affect.
- Human Recognition-Empathic (HR-E), in which a human observer recognized user's affect and prompted the robot (optimal condition).

Our results indicated that the subjects rated empathy and likability of both the AR-E and HR-E robots significantly higher than the non-empathic robot (the control condition). Also, there was no significant difference between subjects' ratings of the task in all conditions. In other words, subjects rated the task engagement similar in all conditions, regardless of being empathized or not. Of course, this finding depends on the type of task. In our experiment, the task (describing videos in one word) was fairly easy and short.

We also calculated the accuracy of the integrated automated FER system on the robot when interacting with different users. Despite the automated FER system was trained on AffectNet which covers a variety of scene lightings, camera views, backgrounds, subjects head-pose and ethnicity, etc., the automated FER system only was 49% correct in our setting. We processed user's facial expressions frame by frame and used a weighted majority voting to determine the users' affect during watching the video clips. Perhaps, a better approach considering spatiotemporal in recognizing facial expressions would result in a better recognition. Surprisingly, despite the lower accuracy of the automated FER, there was no significant difference between subjects' ratings of empathy and likability of AR-E and HR-E (optimal condition). We believe that "self-correction" played an important role in compensating poor accuracy of automated FER system. In other word, subjects preferred a robot with an ability to understand their affect, even with a low accuracy.

Chapter 6

Conclusion and Future Work

This dissertation described our current progress towards designing, manufacturing, and evaluating a perceptive and expressive life-like robotic head, called ExpressionBot. ExpressionBot consists of a neck system and a light projector that projects a facial animation on a 3D translucent facial mask. Since a computer-generated facial animation is projected onto a mask, the rear-projected robotic platform can portray natural and realistic facial movement, and the robotic face can range from cartoon-like to photo-realistic characters. The proposed robotic system, relative to mechatronic and android faces, is thus a highly flexible research tool, mechanically simple, and low-cost to design, build, and maintain (the cost of the hardware system is about $1500).

At first, individuals' experiences of interpreting the facial expressions and the proposed visual speech of ExpressionBot was compared with the facial animation on the computer screen. The results of our initial HRI studies illustrated the benefits and the value of the proposed robotic platform over the same animation displayed on a computer flat screen. The studies indicated that although having embodiment through the rear-projected system does not play any role in improving the perception of visual speech, it can improve the perception of emotive agents in certain emotions such as Anger and Sadness, and highly affects the precision of the eye gaze.

During these experiments, the users were in front of the robot, and it was not clear whether the users benefited from the physicality of the robot or they were under the impression of its physical presence. We then distinguished the role of the robot's embodiment from its physical presence in three facial cues (i.e., visual speech, facial expressions and eye gaze) using a quantitative approach. In particular, three different conditions (i.e., co-present of the robot, telepresent of the robot, and virtual agent) were studied to discover whether the embodiment of the robot has any interaction value proposition compared with an on-screen animation. The results of this study indicate that:

1. Neither embodiment nor presence plays a role in improving the perception of visual speech, regardless of syntactic or semantic cues in sentences.

2. Both embodiment and physical presence improve the perception of certain facial expressions in emotive agents.

3. The combination of embodiment and presence (and mainly embodiment) highly affects the precision of eye gaze perception in a frontal situated setting.

Since, the eventual goal of the ExpressionBot is to interact with users in an uncontrolled setting (aka *"in the wild"* setting), where there is a high variation in scene lighting, camera view, image resolution, background, subjects' head-pose and ethnicity, and existing facial expression recognition systems lack enough generality in the wild, we proposed a new Deep Neural Network (DNN) architecture and created a new database of facial faces with expressions (called **AffectNet**). AffectNet contains more than 1M images with faces and 440,000 manually annotated images with facial expressions and valence and arousal emotions. Only, the process of annotating took more than a year and AffectNet is by far the largest database of facial affect in still images which covers both categorical and dimensional models. Experimental results of the proposed deep neural network architecture and two simple baselines on AffectNet indicated that proposed affect perception system is more accurate than existing expression recognition systems.

We then integrated this automated FER system into the spoken dialog of our robotic platform to extend and enrich the capabilities of ExpressionBot beyond spoken dialog and create an affect-aware robotic agent that can measure and infer users' affect and cognition. We evaluated whether this integration can improve social/interaction aspects of our agent with users. We designed a series of HRI experiments, in which the subjects watched some videos to evoke their emotions and the robot asked them to describe each video in one word. We studied three conditions as: 1) Non-Empathic robot (NE), in which the robot did not recognize user's affect and only performed the task (control condition), 2) Automated Recognition-Empathic (AR-E), in which the developed automated FER system was used to recognize subject's affect, and 3) Human Recognition-Empathic (HR-E), in which a human observer recognized user's affect and prompted the robot (optimal condition). Three social/interaction aspects (task engagement, being empathic, and likability of the robot) were measured in the experiment. Our results indicated that the subjects rated empathy and likability of both the AR-E and HR-E robots significantly higher than the non-empathic robot (the control condition). Also, users rated our affect-aware agent as empathic and likable as a robot in which user's affect is recognized by a human (HR-E).

The developed robotic head represents a new level of integration of emotive capabilities that enables researchers to study socially emotive robots/agents that can generate spoken-language, show emotions, measure and infer users' affect and cognition, and communicate effectively with people in a natural way as humans do. Such systems can be applied in many domains including health-care, education, entertainment, and home-care. It will also be an ideal platform for designing a new generation of more immersive and effective intelligent tutoring and therapy systems, and robot-assisted therapeutic treatments.

While this dissertation has demonstrated the potential of the proposed affect-aware robotic platform, there are some future research and improvements that can be made to enhance the proposed platform. Some of these directions are:

1. **Affect recognition from other modalities:** In this dissertation, we only used facial expression recognition to understand user's affect. Although facial expressions play a vital role in social interaction and they are a common nonverbal channel through which HMI systems can recognize humans' internal emotions, human affect sensing can be obtained from a broader range of behavioral cues and signals such as body gestures, head movements, speech acoustic analysis, dialog sentiment analysis. Using multi-modal affect recognition with audiovisual affect sensing and tactile sensors (e.g., heart rate, skin conductivity, thermal signals etc.) can enable social robots to understand non-visible user's affect beyond basic expressions with higher accuracy.

2. **Considering spatio-temporal features in recognizing facial expressions:** The proposed FER system in this dissertation showed a great accuracy in classifying facial images. However, in practice, the robot interact with users on a continuous basis. Methods such processing user's facial image frame by frame and using majority voting to determine the affect over a period of time can have low accuracy, especially in subtle emotions. An alternative approach is to consider spatio-temporal information to determine user's affect over a period of time. This can be solved by using Recurrent Neural Networks, however, it necessitates creating a large database of videos captured in wild setting.

3. **Improving the facial animation system:** Our comprehensive HRI study on distinguishing the role of embodiment from physical presence, showed that there was not

a significant difference between some emotions and perception of visual speech. The perception of these facial cues highly depends on the algorithm and the models that are used to generate them in the animation. Some of the facial models need to be re-designed by the graphic artist to portray better emotions. Also, the visual speech algorithm has room for improvement as the ground-truth (video of the human) had significantly higher audio-visual intelligibility than our animation/robot. We used a multi-target morphing method to generate the visual speech. Other approaches such as using motion capture to train a visual speech generation system can enhance the performance of visual speech in our animation.

4. **Integrating eye-gaze with spoken dialog:** In this dissertation, we mainly focused on understanding user affect from facial expressions. Eye gaze is also one of the most basic and important features of the human face for nonverbal communication. Considering the importance of eye gaze in social interaction, the agent should be endowed with the ability to perceive user's eye-gaze direction for capturing user's attention and maintaining engagement with the user. In addition, the agent can be empowered with the ability to control its eye-gaze direction during interaction with the user. This can improve the ability to convey information about the emotional and mental state of the agent and makes the interaction more natural.

In summary, this dissertation presented the development and HRI studies of a perceptive, and expressive, conversational, rear-projected, life-like robotic agent (aka ExpressionBot or Ryan) that models natural face-to-face communication between human and emapthic agent. The results of our in-depth and comprehensive human-robot-interaction studies show that this robotic agent can serve as a model for creating the next generation of empathic social robotic agents for a variety of applications including but not limit ted to aging health-care, education, and entertainment.

Bibliography

S. Al Moubayed and G. Skantze. Perception of gaze direction for situated interaction. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, page 3. ACM, 2012. 51, 52

S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström. Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems*, pages 114–130. Springer, 2012. 9, 24, 34

S. Al Moubayed, G. Skantze, and J. Beskow. The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics*, 10(01):1350005, 2013. 28, 34, 37, 64, 65

Albert-Hubo. Albert einstein hubo, 2005. URL `http://www.hansonrobotics.com/robot/albert-einstein-hubo/`. 8

T. Allison, A. Puce, and G. McCarthy. Social perception from visual cues: role of the sts region. *Trends in cognitive sciences*, 4(7):267–278, 2000. 51

S. M. Anstis, J. W. Mayhew, and T. Morley. The perception of where a face or television'portrait'is looking. *The American journal of psychology*, 82(4):474–489, 1969. 51

S. Arora, A. Bhaskara, R. Ge, and T. Ma. Provable bounds for learning some deep representations. *arXiv preprint arXiv:1310.6343*, 2013. 94

I. P. Association. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999. 12

W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati. The benefits of interactions with physically present robots over video-displayed agents. *International Journal of Social Robotics*, 3(1):41–52, 2011. 28

S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International journal of computer vision*, 56(3):221–255, 2004. 78, 79, 81, 91

E. Bal, E. Harden, D. Lamb, A. Van Hecke, J. Denver, and S. Porges. Emotion recognition in children with autism spectrum disorders: Relations to eye gaze and autonomic state. *Journal of Autism and Developmental Disorders*, 40(3):358–370, 2010. ISSN 0162-3257. doi: 10.1007/s10803-009-0884-3. URL `http://dx.doi.org/10.1007/s10803-009-0884-3`. 94

T. Bänziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook*, pages 271–294, 2010. 99

S. Baron-Cohen, R. Campbell, A. Karmiloff-Smith, J. Grant, and J. Walker. Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology*, 13(4):379–398, 1995. 51

C. Bartneck, J. Reichenbach, and v. A. Breemen. In your face, robot! the influence of a character's embodiment on how users perceive its emotional expressions. In *Proceedings of the Design and Emotion*, pages 32–51, 2004. 41, 42, 65

C. Becker-Asano and H. Ishiguro. Evaluating facial displays of emotion for the android robot geminoid f. In *2011 IEEE Workshop on Affective Computational Intelligence (WACI)*, pages 1–8, April 2011. doi: 10.1109/WACI.2011.5953147. 40

J. M. Beer, A. Prakash, T. L. Mitzner, and W. A. Rogers. Understanding robot acceptance. *Georgia Institute of Technology*, pages 1–45, 2011. 40

C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR'16), Las Vegas, NV, USA*, 2016. 105, 109, 112, 117

S. Bermejo and J. Cabestany. Oriented principal component analysis for large margin classifiers. *Neural Networks*, 14(10):1447–1461, 2001. 113

T. W. Bickmore and R. W. Picard. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327, 2005. 143

R. C. Bilger, J. Nuetzel, W. Rabinowitz, and C. Rzeczkowski. Standardization of a test of speech perception in noise. *Journal of Speech, Language, and Hearing Research*, 27(1): 32–48, 1984. 37

F. Biocca. The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication*, 3(2):0–0, 1997. 29

D. Bolanos. The bavieca open-source speech recognition toolkit. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 354–359. IEEE, 2012. 11, 36, 38

M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1): 49–59, 1994. 111

S. Brave, C. Nass, and K. Hutchinson. Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, 62(2):161–178, 2005. 143

C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1):119–155, 2003. 40

C. Breazeal. Socially intelligent robots. *interactions*, 12(2):19–22, 2005. 1

C. Breazeal, K. Dautenhahn, and T. Kanda. Social robotics. In *Springer handbook of robotics*, pages 1935–1972. Springer, 2016. 1

C. L. Breazeal. *Sociable machines: Expressive social exchange between humans and robots*. PhD thesis, Massachusetts Institute of Technology, 2000. 33

A. Bruce, I. Nourbakhsh, and R. Simmons. The role of expressiveness and attention in human-robot interaction. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 4, pages 4138–4142. IEEE, 2002. 40

G. Caridakis, K. Karpouzis, and S. Kollias. User and context adaptive neural networks for emotion recognition. *Neurocomputing*, 71(13):2553–2562, 2008. 128

J. Cassell. *Embodied conversational agents*. MIT press, 2000. 26, 29, 40, 50

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 149

Y.-C. Chen and S.-L. Yeh. Look into my eyes and i will see you: Unconscious processing of human gaze. *Consciousness and cognition*, 21(4):1703–1710, 2012. 49

M. G. Cline. The perception of where a person is looking. *The American journal of psychology*, 80(1):41–50, 1967. 53

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37, 1960. 113, 128

J. Cohen. Statistical power analysis for the behavioral sciences (revised ed.), 1977. 49, 159

J. F. Cohn, Z. Ambadar, and P. Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, pages 203–221, 2007. 73

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001. 77, 78

T. F. Cootes, C. J. Taylor, et al. Statistical models of appearance for computer vision, 2004. 78, 80, 86

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 127, 133

H. Cramer, J. Goddijn, B. Wielinga, and V. Evers. Effects of (in) accurate empathy and situational valence on attitudes towards robots. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, pages 141–142. IEEE, 2010. 145

L. J. Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3): 297–334, 1951. 27

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 133, 135

K. Dautenhahn. The art of designing socially intelligent agents: Science, fiction, and the human in the loop. *Applied artificial intelligence*, 12(7-8):573–617, 1998. 29

K. Dautenhahn. Socially intelligent agents-the human in the loop. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(5):345–348, 2001. 29

K. Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, 2007. 1

K. Dautenhahn, B. Ogden, and T. Quick. From embodied to socially embedded agents–implications for interaction-aware robots. *Cognitive Systems Research*, 3(3):397–428, 2002. 21

M. H. Davis. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113, 1983. 156

M. H. Davis. *Empathy: A social psychological approach.* Westview Press, 1994. 141

F. De la Torre and J. F. Cohn. Facial expression analysis. In *Visual analysis of humans*, pages 377–409. Springer, 2011. 73

F. Delaunay, J. de Greeff, and T. Belpaeme. A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 39–44. IEEE Press, 2010. 28, 51, 52, 67

F. Delaunay, J. de Greeff, and T. Belpaeme. Lighthead robotic face. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*, pages 101–101, 2011. 9

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 98

A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2106–2112. IEEE, 2011. 99, 107

A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge 2013. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 509–516. ACM, 2013. 70, 77, 105, 107, 112

DreamFace-Tech. Social robotics, 2015. URL `http://dreamfacetech.com/`. last checked: 01.20.2017. 19

S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014. 104, 109

P. Ekman and W. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto, 1978. 43

P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971. 70, 104

P. Ekman and W. V. Friesen. Facial action coding system. *Consulting Psychologists Press*, 1977. 14, 73, 104

S. Eleftheriadis, O. Rudovic, and M. Pantic. Discriminative shared gaussian processes for multi-view and view-invariant facial expression recognition. *IEEE Transactions on Image Processing*, 24(1):189–204, 2015. 100

L. L. Elliott. Verbal auditory closure and the speech perception in noise (spin) test. *Journal of Speech, Language, and Hearing Research*, 38(6):1363–1376, 1995. 37

N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604, 2000. 50

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969. 13

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008. 133, 136

Y. Fan, X. Lu, D. Li, and Y. Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM, 2016. 115, 117

J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007. 105

W. V. Friesen and P. Ekman. Emfacs-7: Emotional facial action coding system. *Unpublished manuscript, University of California at San Francisco*, 2:36, 1983. 44, 74, 104

R. Fujimura, K. Nakadai, M. Imai, and R. Ohmura. Prot - an embodied agent for intelligible and user-friendly human-robot interaction. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 3860–3867. IEEE, 2010. 28, 29

Geminoid. Hiroshi ishiguro laboratories, 2011. URL `http://www.geminoid.jp/en/robots.html`. 1, 8

R. Gockley, R. Simmons, J. Wang, D. Busquets, C. DiSalvo, K. Caffrey, S. Rosenthal, J. Mink, S. Thomas, and W. Adams. Grace and george: Social robots at AAAI. In *Proceedings of AAAI*, volume 4, pages 15–20, 2004. 8

E. Goffman. Behavior in public place. *Glencoe: the free press, New York*, 1963. 30

I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64:59–63, 2015. 98, 99, 105, 108, 112

R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005. 78, 81, 82, 86, 92

R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 70, 86, 87, 98, 99, 107, 112

E. Guizzo. World robot population reaches 8.6 million. *IEEE Spectrum*, 14, 2010. 1, 25

A. Hartholt, J. Gratch, L. Weiss, et al. At the virtual frontier: Introducing gunslinger, a multi-character, mixed-reality, story-driven experience. In *International Workshop on Intelligent Virtual Agents*, pages 500–501. Springer, 2009. 26

B. Hasani and M. H. Mahoor. Facial affect estimation in the wild using deep residual and convolutional networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 1955–1962. IEEE, 2017. 149

M. Hashimoto and D. Morooka. Facial expression of a robot using a curved surface display. In *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 765–770, 2005. 9

B. Hayes, D. Ullman, E. Alexander, C. Bank, and B. Scassellati. People help robots who help others, not robots who help themselves. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 255–260. IEEE, 2014. 143

H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 128

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 149, 150

L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. ACM, 2015. 115, 116, 117

S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017. 149

M. L. Hoffman. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, 2001. 141

M. E. Hoque, M. Courgeon, J.-C. Martin, B. Mutlu, and R. W. Picard. Mach: My automated conversation coach. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 697–706. ACM, 2013. 26

HRP-4C. Hrp-4c, 2009. URL `https://en.wikipedia.org/wiki/HRP-4C`. 8

IDC. International data corporation (idc) press., 2016. URL `http://www.idc.com/getdoc.jsp?containerId=prUS41046916`. last checked: 05.14.2017. 1, 25

M. Imai, T. Kanda, T. Ono, H. Ishiguro, and K. Mase. Robot mediated round table: Analysis of the effect of robot's gaze. In *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*, pages 411–416. IEEE, 2002. 50

I. P. A. IPA-Handbook. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999. 36

R. J. Itier and M. Batty. Neural bases of eye and gaze processing: the core of social cognition. *Neuroscience & Biobehavioral Reviews*, 33(6):843–863, 2009. 51

L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 245–251. IEEE, 2013. 113, 128

Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 100, 130

W. Ju and D. Sirkin. Animate objects: How physical motion encourages public interaction. In *International Conference on Persuasive Technology*, pages 40–51. Springer, 2010. 26, 28, 29

S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari, et al. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM, 2013. 71, 77

S. Kajita, T. Nakano, M. Goto, Y. Matsusaka, S. Nakaoka, and K. Yokoi. Vocawatcher: Natural singing motion generator for a humanoid robot. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 2000–2007. IEEE, 2011. 33

D. N. Kalikow, K. N. Stevens, and L. L. Elliott. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America*, 61(5):1337–1351, 1977. 37

S.-H. Kang, J. H. Watt, and S. K. Ala. Communicators' perceptions of social presence as a function of avatar realism in small display mobile communication devices. In *Hawaii*

International Conference on System Sciences, Proceedings of the 41st Annual, pages 147–147. IEEE, 2008. 145

J. Kätsyri and M. Sams. The effect of dynamics on identifying basic emotions from synthetic and natural faces. *International Journal of Human-Computer Studies*, 66(4):233–242, 2008. 41, 42

Y. Keller and A. Averbuch. Fast motion estimation using bidirectional gradient methods. *IEEE Transactions on Image Processing*, 13(8):1042–1054, 2004. 79

A. Kendon. Some functions of gaze-direction in social interaction. *Acta psychologica*, 26: 22–63, 1967. 50

A. Kendon, R. M. Harris, and M. R. Key. *Organization of behavior in face-to-face interaction*. Walter de Gruyter, 1975. 27

M. Kenji. Recognition of facial expression from optical flow. *IEICE TRANSACTIONS on Information and Systems*, 74(10):3474–3483, 1991. 73

C. D. Kidd and C. Breazeal. Effect of a robot on user perceptions. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 4, pages 3559–3564. IEEE, 2004. 26, 28, 29

S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition*, 26(2):169–181, 2008. 27, 28, 29

M. Kinya and E. Mitsuo. Illusory face dislocation effect and configurational integration in the inverted face. *Tohoku Psychologica Folia*, 43(1-4):150–160, 1984. 53

Kismet. Kismet, 2002. URL `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html`. last checked: 01.20.2017. 1, 7

N. L. Kluttz, B. R. Mayes, R. W. West, and D. S. Kerby. The effect of head turn on the perception of gaze. *Vision research*, 49(15):1979–1993, 2009. 53

H. Kobayashi and F. Hara. Facial interaction between animated 3d face robot and human beings. In *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, volume 4, pages 3732–3737. IEEE, 1997. 73, 115

S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. 110, 112, 117

S. Kopp, L. Gesellensetter, N. C. Krämer, and I. Wachsmuth. A conversational agent as museum guide–design and evaluation of a real-world application. In *International Workshop on Intelligent Virtual Agents*, pages 329–343. Springer, 2005. 26

H. Kose-Bagci, E. Ferrari, K. Dautenhahn, D. S. Syrdal, and C. L. Nehaniv. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics*, 23(14):1951–1996, 2009. 28

K. Krippendorff. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30(1):61–70, 1970. 113, 128

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 71, 76, 95, 115, 128, 134

T. Kuratate, Y. Matsusaka, B. Pierce, and G. Cheng. Mask-bot: a life-size robot head using talking head animation for human-robot communication. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, pages 99–104, 2011. 9

S. S. Kwak, Y. Kim, E. Kim, C. Shin, and K. Cho. What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot. In *RO-MAN, 2013 IEEE*, pages 180–185. IEEE, 2013. 143

S. R. Langton, H. Honeyman, and E. Tessler. The influence of head contour and nose angle on the perception of eye-gaze direction. *Perception & psychophysics*, 66(5):752–771, 2004. 53

N. Lazzeri, D. Mazzei, A. Greco, A. Rotesi, A. Lanatà, and D. E. De Rossi. Can a humanoid face be expressive? a psychophysiological investigation. *Frontiers in bioengineering and biotechnology*, 3, 2015. 42, 43, 66

K. M. Lee, Y. Jung, J. Kim, and S. R. Kim. Are physically embodied social agents better than disembodied social agents?: The effects of physical embodiment, tactile interaction, and people's loneliness in human–robot interaction. *International Journal of Human-Computer Studies*, 64(10):962–973, 2006. 28

S. H. Lee, K. Plataniotis, and Y. M. Ro. Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing*, page 1, 2014. 100

I. Leite, C. Martinho, A. Pereira, and A. Paiva. As time goes by: Long-term evaluation of social presence in robotic companions. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pages 669–674. IEEE, 2009. 146

I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. Empathic robots for long-term interaction. *International Journal of Social Robotics*, 6(3):329–341, 2014. 146

J. C. Lester, J. L. Voerman, S. G. Towns, and C. B. Callaway. Deictic believability: Coordinated gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13(4-5):383–414, 1999. 33

D. Li, P. P. Rau, and Y. Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2(2):175–186, 2010. 157

J. Li. The benefit of being physically present: a survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies*, 77:23–37, 2015. 27, 28, 30

J. Li and M. Chignell. Communication of emotion in social robots through simple head and arm movements. *International Journal of Social Robotics*, 3(2):125–142, 2011. 28

C.-Y. Lin, L.-C. Cheng, and L.-C. Shen. Oral mechanism design on face robot for lip-synchronized speech. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 4316–4321. IEEE, 2013a. 33

M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013b. 75, 94

P. Lincoln, G. Welch, A. Nashel, A. Ilie, and H. Fuchs. Animatronic shader lamps avatars. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 27–33. IEEE, 2009. 9

C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *Image processing, IEEE Transactions on*, 11 (4):467–476, 2002. 73, 115

M. Liu, S. Li, S. Shan, and X. Chen. Au-aware deep networks for facial expression recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013. 76, 100

M. Liu, S. Li, S. Shan, R. Wang, and X. Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *Computer Vision–ACCV 2014*, pages 143–157. Springer, 2014a. 77, 100

M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM, 2014b. 77

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 94–101. IEEE, 2010. 45, 73, 99, 107, 109, 112

P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011. 109

R. C. Luo, S.-R. Chang, C.-C. Huang, and Y.-P. Yang. Human robot interactions using speech synthesis and recognition with lip synchronization. In *IECON 2011-37th Annual Conference on IEEE Industrial Electronics Society*, pages 171–176. IEEE, 2011. 33

M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205. IEEE, 1998. 70, 107

J. Ma and R. Cole. Animating visible speech and facial expressions. *The Visual Computer*, 20(2-3):86–105, 2004. 11, 36

M. H. Mahoor, M. Abdel-Mottaleb, A.-N. Ansari, et al. Improved active shape model for facial feature extraction in color images. *Journal of multimedia*, 1(4):21–28, 2006. 85

M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 74–80. IEEE, 2009. 77

M. H. Mahoor, A. Mollahosseini, L. Skelly, and Q. Martindale-George. Rear-projected life-like robotic head, Feb 2015. US Patent App. 15/042,002. 18

S. C. Marsella, W. L. Johnson, and C. LaBore. Interactive pedagogical drama. In *Proceedings of the fourth international conference on Autonomous agents*, pages 301–308. ACM, 2000. 143

S. Mascarenhas, R. Prada, A. Paiva, and G. J. Hofstede. Social importance dynamics: A model for culturally-adaptive agents. In *International Workshop on Intelligent Virtual Agents*, pages 325–338. Springer, 2013. 143

I. Matthews and S. Baker. Active appearance models revisited. *International journal of computer vision*, 60(2):135–164, 2004. 77, 78, 81, 86

S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *Affective Computing, IEEE Transactions on*, 4(2):151–160, 2013. 70, 73, 99, 107, 109, 112, 115

C. Mayer, M. Eggers, and B. Radig. Cross-database evaluation for facial expression recognition. *Pattern recognition and image analysis*, 24(1):124–132, 2014. 71, 100, 102, 103

D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013. 107, 108, 109, 112, 117

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976. 32

G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012. 111

R. Mégret, J.-B. Authesserre, and Y. Berthoumieu. Bidirectional composition on lie groups for gradient-based image alignment. *IEEE Transactions on Image Processing*, 19(9): 2369–2381, 2010. 79

K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, volume 964, pages 965–966. Citeseer, 1999. 107

Y.-Q. Miao, R. Araujo, and M. S. Kamel. Cross-domain facial expression recognition using supervised kernel mean matching. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 326–332. IEEE, 2012. 102

Microsoft-Oxford. Microsoft cognitive services - emotion api, 2015. URL `https://www.microsoft.com/cognitive-services/en-us/emotion-api`. (Accessed on 12/01/2016). 134

P. Milgram, H. Takemura, A. Utsumi, and F. Kishino. Augmented reality: A class of displays on the reality-virtuality continuum. In *Photonics for industrial applications*, pages 282–292. International Society for Optics and Photonics, 1995. 30

S. Milgram. Behavioral study of obedience. *The Journal of abnormal and social psychology*, 67(4):371, 1963. 143

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995. 109

K. Misawa, Y. Ishiguro, and J. Rekimoto. Livemask: A telepresence surrogate system with a face-shaped screen for supporting nonverbal communication. In *Proceedings of the international working conference on advanced visual interfaces*, pages 394–397. ACM, 2012. 51, 52, 67

M. Mohammadi, E. Fatemizadeh, and M. Mahoor. Pca-based dictionary building for accurate facial expression recognition via sparse representation. *Journal of Visual Communication and Image Representation*, 25(5):1082 – 1092, 2014. ISSN 1047-3203. doi: http:

//dx.doi.org/10.1016/j.jvcir.2014.03.006. URL `http://www.sciencedirect.com/science/article/pii/S1047320314000625`. 73, 115

M. R. Mohammadi, E. Fatemizadeh, and M. H. Mahoor. Intensity estimation of spontaneous facial action units based on their sparsity properties. *IEEE transactions on cybernetics*, 46(3):817–826, 2016. 70

A. Mollahosseini and M. H. Mahoor. Bidirectional warping of active appearance model. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 875–880. IEEE, 2013. 95, 108

A. Mollahosseini, G. Graitzer, E. Borts, S. Conyers, R. M. Voyles, R. Cole, and M. H. Mahoor. Expressionbot: An emotive lifelike robotic face for face-to-face communication. In *Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on*, pages 1098–1103. IEEE, 2014. 28, 34, 35, 37, 42, 44, 51, 52, 64, 65, 67, 70

A. Mollahosseini, D. Chan, and M. H. Mahoor. Going deeper in facial expression recognition using deep neural networks. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016a. 115, 117

A. Mollahosseini, B. Hasani, M. J. Salvador, H. Abdollahi, D. Chan, and M. H. Mahoor. Facial expression recognition from world wild web. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2016b. 105, 108, 112, 115, 117

M. Mori, K. MacDorman, and N. Kageki. The uncanny valley [from the field]. *Robotics Automation Magazine, IEEE*, 19(2):98–100, June 2012. ISSN 1070-9932. doi: 10.1109/MRA.2012.2192811. 2

S. A. Moubayed, J. Edlund, and J. Beskow. Taming mona lisa: Communicating gaze faithfully in 2d and 3d facial projections. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 1(2):11, 2012. 51, 52, 67

B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 61–68. ACM, 2009. 50

M. A. Nicolaou, H. Gunes, and M. Pantic. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3695–3699. IEEE, 2010. 112

M. A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105, 2011. 112, 113, 114, 117

K.-G. Oh, C.-Y. Jung, Y.-G. Lee, and S.-J. Kim. Real-time lip synchronization between text-to-speech (tts) system and robot mouth. In *RO-MAN, 2010 IEEE*, pages 620–625. IEEE, 2010. 33

B. L. Omdahl. Cognitive appraisal, emotion, and empathy. *Psychology Press*, 1995. 141

Y. Otsuka, I. Mareschal, A. J. Calder, and C. W. Clifford. Dual-route model of the effect of head orientation on perceived gaze direction. *Journal of Experimental Psychology: Human perception and performance*, 40(4):1425, 2014. 53

S. Ouni, D. W. Massaro, M. M. Cohen, K. Young, and A. Jesse. Internationalization of a talking head. In *Proc. of 15th International Congress of Phonetic Sciences, Barcelona, Spain*, pages 286–318, 2003. 33, 34

A. Paiva, J. Dias, D. Sobral, R. Aylett, P. Sobreperez, S. Woods, C. Zoll, and L. Hall. Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 194–201. IEEE Computer Society, 2004. 143

A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):11, 2017. 142, 143, 144, 145, 156

G. Paltoglou and M. Thelwall. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing*, 4(1):116–123, 2013. 123

M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005. 70, 99, 107, 112

Paro. Paro therapeutic robot, 2004. URL `http://www.parorobots.com/`. 1

N. Pateromichelakis, A. Mazel, M. Hache, T. Koumpogiannis, R. Gelin, B. Maisonnier, and A. Berthoz. Head-eyes system and gaze analysis of the humanoid robot romeo. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 1374–1379. IEEE, 2014. 50

D. Perrett, P. Smith, D. Potter, A. Mistlin, A. Head, A. Milner, and M. Jeeves. Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London B: Biological Sciences*, 223(1232):293–317, 1985. 50

R. Pfeifer and C. Scheier. *Understanding intelligence*. MIT press, 1999. 29

B. Pierce, T. Kuratate, A. Maejima, S. Morishima, Y. Matsusaka, M. Durkovic, K. Diepold, and G. Cheng. Development of an integrated multi-modal communication robotic face. In *Advanced Robotics and its Social Impacts (ARSO), 2012 IEEE Workshop on*, pages 101–102, 2012. 9

J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun. Towards robotic assistants in nursing homes: Challenges and results. *Robotics and autonomous systems*, 42(3-4): 271–281, 2003. 2

H. Prendinger and M. Ishizuka. The empathic companion: A character-based interface that addresses users'affective states. *Applied Artificial Intelligence*, 19(3-4):267–285, 2005. 143

S. D. Preston and F. B. De Waal. Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, 25(1):1–20, 2002. 141, 150

T. E. Project. The emote project, 2013. URL `http://gaips.inesc-id.pt/emote/`. last checked: 01.20.2017. 156

P. P. Rau, Y. Li, and D. Li. Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior*, 25(2):587–595, 2009. 157

S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014. 108, 119, 149

S. Reysen. Construction of a new scale: The reysen likability scale. *Social Behavior and Personality: an international journal*, 33(2):201–208, 2005. 157

L. D. Riek and P. Robinson. Using robots to help people habituate to visible disabilities. In *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, pages 1–8. IEEE, 2011. 157

L. D. Riek, P. C. Paul, and P. Robinson. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. *Journal on Multimodal User Interfaces*, 3(1-2):99–108, 2010. 145

F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013. 109, 110, 112

F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1335–1336. ACM, 2015. 111, 113

A. M. Rosenthal-von der Pütten, N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler. An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, 5(1):17–34, 2013. 143

K. Ruhland, S. Andrist, J. Badler, C. Peters, N. Badler, M. Gleicher, B. Mutlu, and R. Mcdonnell. Look me in the eyes: A survey of eye and gaze animation for virtual agents and artificial systems. In *Eurographics State-of-the-Art Report*, pages 69–91, 2014. 50

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014. 75, 94

J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39 (6):1161–1178, 1980. 104, 105, 123

C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 896–903, 2013. 119

C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: Database and results. *Image and Vision Computing*, 47:3–18, 2016. 119

A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008. 43

B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, pages 415–424. Springer, 2011. 111, 113

B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456. ACM, 2012. 111, 113

Second-Life. Second life official site - virtual worlds, avatars, free 3d chat, 2003. URL `http://secondlife.com/`. 8

S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young. Poor thing! would you feel sorry for a simulated robot?: A comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 125–132. ACM, 2015. 143, 145

C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009. 73, 100, 102, 103, 115, 117

P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979. 113

C. Siciliano, A. Faulkner, and G. Williams. Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*, 2003. 33, 34

A. Smola and V. Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9:155–161, 1997. 127, 135

I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, 3(1):32–41, 2012. 107, 109, 110

M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian Joint Conference on Artificial Intelligence*, pages 1015–1021. Springer, 2006. 113

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 76

E. Staub. Commentary on part i. *Empathy and its development*, pages 103–115, 1987. 141

Y. Sun, D. Liang, X. Wang, and X. Tang. Deepid3: Face recognition with very deep neural networks. *CoRR*, abs/1502.00873, 2015. URL `http://arxiv.org/abs/1502.00873`. 94

J. M. Susskind, A. K. Anderson, and G. E. Hinton. The toronto face database. *Technical report, UTML TR 2010-001, University of Toronto*, 2010. 77

T. D. Sweeny and D. Whitney. The center of attention: Metamers, sensitivity, and bias in the emergent perception of gaze. *Vision Research*, 131:67–74, 2017. 53

T. D. Sweeny, E. Guzman-Martinez, L. Ortega, M. Grabowecky, and S. Suzuki. Sounds exaggerate visual shape. *Cognition*, 124(2):194–200, 2012a. 33

T. D. Sweeny, S. Haroz, and D. Whitney. Reference repulsion in the categorical perception of biological motion. *Vision research*, 64:26–34, 2012b. 51

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014. 71, 75, 94

S. Taheri, Q. Qiang, and R. Chellappa. Structure-preserving sparse decomposition for facial expression analysis. *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society*, 23(8):3590, 2014. 74, 100

Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014. 71, 115

Y. Tang. Deep learning using linear support vector machines. *arXiv preprint arXiv:1306.0239*, 2013. 100, 115, 117

A. Tapus and M. J. Mataric. Emulating empathy in socially assistive robotics. In *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics*, pages 93–96, 2007. 156

Y.-I. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001. 107

D. Todorović. Geometrical basis of perception of gaze direction. *Vision research*, 46(21): 3549–3562, 2006. 8, 21, 50

A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1653–1660. IEEE, 2014. 71, 115

UCR-team. Facial expression recognition and analysis challenge (fera2011), 2011. URL `http://sspnet.eu/fera2011/`. (Accessed on 12/01/2016). 100

M. Vala, P. Sequeira, A. Paiva, and R. Aylett. Fearnot! demo: a virtual environment with synthetic characters to help bullying. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, page 271. ACM, 2007. 26

M. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection in video. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 635–640. IEEE, 2004. 73

M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013. 111

M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014. 111

M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016-depression, mood, and emotion recognition workshop and challenge. *arXiv preprint arXiv:1605.01600*, 2016. 111, 113

M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926. IEEE, 2011. 74, 100

A. Van-Breemen. Bringing robots to life: Applying principles of animation to robots. In *Proceedings of Shapping Human-Robot Interaction workshop held at CHI 2004*, pages 143–144. Citeseer, 2004. 40, 41, 146

A. van Breemen, X. Yan, and B. Meerbeek. icat: an animated user-interface robot with personality. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 143–144. ACM, 2005. 7

L. P. Vardoulakis, L. Ring, B. Barry, C. L. Sidner, and T. Bickmore. Designing relational agents as long term social companions for older adults. In *International Conference on Intelligent Virtual Agents*, pages 289–302. Springer, 2012. 2

P. Viola and M. J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004. 109

J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric. The role of physical embodiment in human-robot interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, pages 117–122. IEEE, 2006. 21

J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric. Embodiment and human-robot interaction: A task-based perspective. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 872–877. IEEE, 2007. 27, 28, 29

J. H. Walker, L. Sproull, and R. Subramani. Using a human face in an interface. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 85–91. ACM, 1994. 26, 33

W. H. Wollaston. On the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London*, 114:247–256, 1824. 53

X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE, 2013. 95

Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto. Responsive robot gaze to interaction partner. In *Robotics: Science and systems*, 2006. 50

D. You, O. C. Hamsici, and A. M. Martinez. Kernel optimization in discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):631–638, 2011. 109

L. Yu. face-alignment-in-3000fps. `https://github.com/yulequan/face-alignment-in-3000fps`, 2016. 108, 119, 149

S. Zafeiriou, A. Papaioannou, I. Kotsia, M. A. Nicolaou, and G. Zhao. Facial affect “in-the-wild": A survey and a new database. In *International Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Affect "in-the-wild" Workshop*, June 2016. 105, 110, 111, 112

Zeno. Zeno, 2009. URL `http://www.hansonrobotics.com/robot/zeno/`. 8

X. Zhang, A. Mollahosseini, B. Kargar, H. Amir, E. Boucher, R. M. Voyles, R. Nielsen, and M. Mahoor. ebear: An expressive bear-like robot. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 969–974. IEEE, 2014. 73

X. Zhang, M. H. Mahoor, and S. M. Mavadati. Facial expression recognition using {l} _ {p}-norm mkl multiclass-svm. *Machine Vision and Applications*, pages 1–17, 2015. 73, 100, 102, 117

G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007. 73

S. Zhao. Toward a taxonomy of copresence. *Presence: Teleoperators and Virtual Environments*, 12(5):445–455, 2003. 30

W. Zhen and Y. Zilu. Facial expression recognition based on local phase quantization and sparse representation. In *Natural Computation (ICNC), 2012 Eighth International Conference on*, pages 222–225. IEEE, 2012. 73, 115

Appendix A

Publications

Journal papers

1. **Mollahosseini, Ali**, Behzad Hasani, and Mohammad H. Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild.” *IEEE Transactions on Affective Computing*, 2017.

2. **Mollahosseini, Ali**, Hojjat Abdollahi, Timothy Sweeny, Ron Cole and Mohammad H. Mahoor. ‘Role of Embodiment and Presence in Human Perception of Robots’ Facial Cues.” *International Journal of Human-Computer Studies* (submitted).

3. Rezaeilouyeh, Hadi, **Ali Mollahosseini**, and Mohammad H. Mahoor. “Microscopic medical image classification framework via deep learning and shearlet transform.” *Journal of Medical Imaging 3*, no. 4, 2016.

Conference papers

1. Abdollahi Hojjat, **Ali Mollahosseini**, Josh T Lane, Mohammad H Mahoor, “A Pilot Study on Using an Intelligent Life-like Robot as a Companion for Elderly Individuals with Dementia and Depression.” *IEEE-RAS International Conference on Humanoid Robots*, IEEE, 2017.

2. **Mollahosseini, Ali**, Behzad Hassani, Michelle J. Salvador, Hojjat Abdollahi, David Chan, and Mohammad H. Mahoor. “Facial Expression Recognition from World Wild Web.” *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016.

3. **Mollahosseini, Ali**, David Chan, and Mohammad H. Mahoor. "Going deeper in facial expression recognition using deep neural networks." *In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1-10. IEEE, 2016.

4. **Mollahosseini, Ali**, Gabriel Graitzer, Eric Borts, Stephen Conyers, Richard M. Voyles, Ronald Cole, and Mohammad H. Mahoor. "ExpressionBot: an emotive lifelike robotic face for face-to-face communication." *In 2014 IEEE-RAS International Conference on Humanoid Robots*, pp. 1098-1103. IEEE, 2014.

5. Zhang, Xiao, **Ali Mollahosseini**, Evan Boucher, Richard M. Voyles, Rodney Nielsen, and Mohammd H. Mahoor. "ebear: An expressive bear-like robot." *In The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 969-974. IEEE, 2014.

6. Kargar, B. Amir H., **Ali Mollahosseini**, Taylor Struemph, Wilson Pace, Rodney D. Nielsen, and Mohammad H. Mahoor. "Automatic measurement of physical mobility in Get-Up-and-Go Test using kinect sensor." *In 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3492-3495. IEEE, 2014.

7. **Mollahosseini, Ali**, and Moohammad Mahoor. "Bidirectional warping of active appearance model." *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 875-880. 2013.

Patents

1. Mahoor, Mohhammad, **Ali Mollahosseini**, Luke Skelly, and Quinn Martindale-George, "Rear-Projected Life-Like Robotic Head." *Patent US-2016-0231645-A1*, 2015.

2. Mahoor, Mohhammad, Hadi Rezaeilouyeh, and **Ali Mollahosseini**. "Methods and Systems for Human Tissue Analysis using Shearlet Transforms." 2016 (under review).

Appendix B

AffectNet

Table B.1: Agreement percentage between two annotators in categorical model of affect (%)

	A1*	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
A1	0.0**	69	70	68	0	0	0	0	0	0	0	0
A2	69	0	64.9	68.3	0	0	0	64.7	0	0	0	0
A3	70	64.9	0	70.6	67.4	69.9	63	62.3	0	48.1	0	0
A4	68	68.3	70.6	0	70.4	70.8	64.3	67.5	0	27.5	0	0
A5	0	0	67.4	70.4	0	70.6	0	0	0	0	0	0
A6	0	0	69.9	70.8	70.6	0	0	0	0	0	0	0
A7	0	0	63	64.3	0	0	0	0	0	75.8	0	0
A8	0	64.7	62.3	67.5	0	0	0	0	51.1	0	0	0
A9	0	0	0	0	0	0	0	51.1	0	0	54.4	0
A10	0	0	48.1	27.5	0	0	75.8	0	0	87.5	0	61.9
A11	0	0	0	0	0	0	0	0	54.4	0	0	0
A12	0	0	0	0	0	0	0	0	0	61.9	0	0

* A1 to A12 indicate Annotators 1 to 12

** Zero means that there were no common images between the two annotators

Table B.2: Samples of annotated categories for queried emotion terms

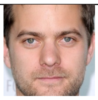
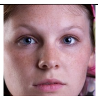
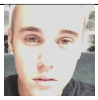
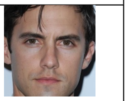
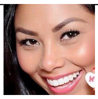
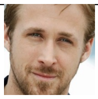
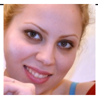
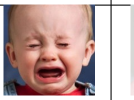
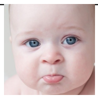
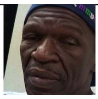
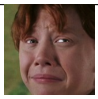
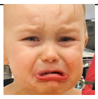
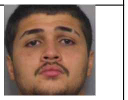
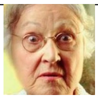
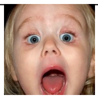
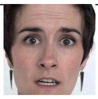
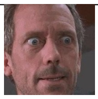
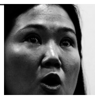
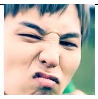
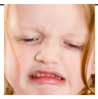
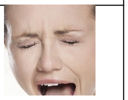
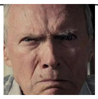
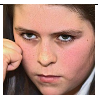
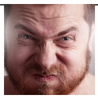
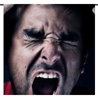
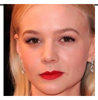
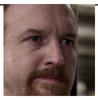
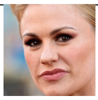
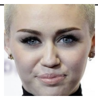
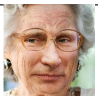
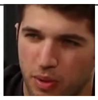
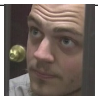
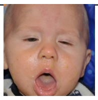
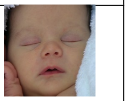
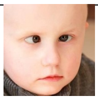
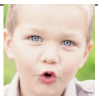
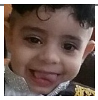
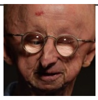
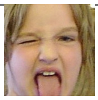
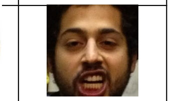
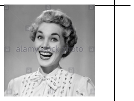
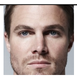
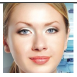
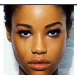
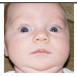
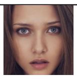
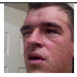
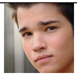
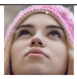
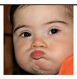
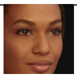
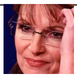
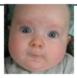
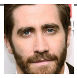
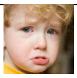
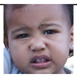
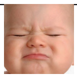
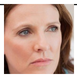
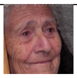
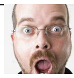
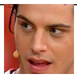
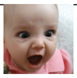
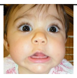
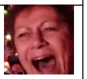
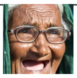
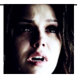
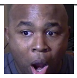
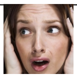
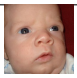
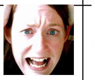
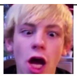
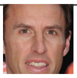
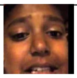
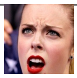
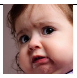
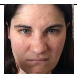
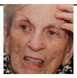
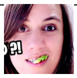
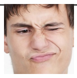
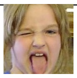
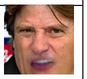
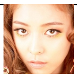
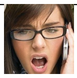
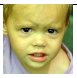
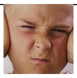
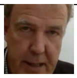
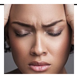
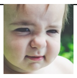
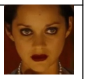
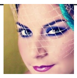
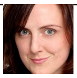
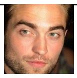
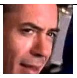
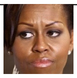
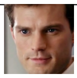
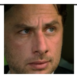
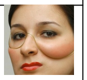
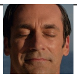
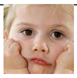
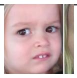
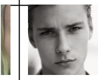
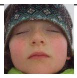
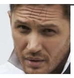
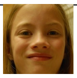
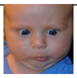
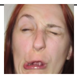
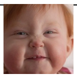
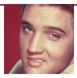
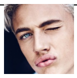
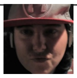
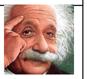
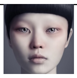
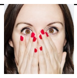
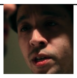
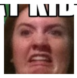
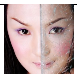
		Queried Expression						
		Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt
Annotated Expression	Neutral							
	Happy							
	Sad							
	Surprise							
	Fear							
	Disgust							
	Anger							
	Contempt							
	None							
	Uncertain							
	Non-Face							

Table B.3: Samples of annotated images by two annotators (randomly selected)

		Annotator 1										
		Neutral	Happy	Sad	Surprise	Fear	Disgust	Anger	Contempt	None	Uncertain	Non-Face
Annotator 2	Neutral											
	Happy											
	Sad											
	Surprise											
	Fear								No Disagreement			
	Disgust											
	Anger											
	Contempt											
	None											
	Uncertain											
	Non-Face											

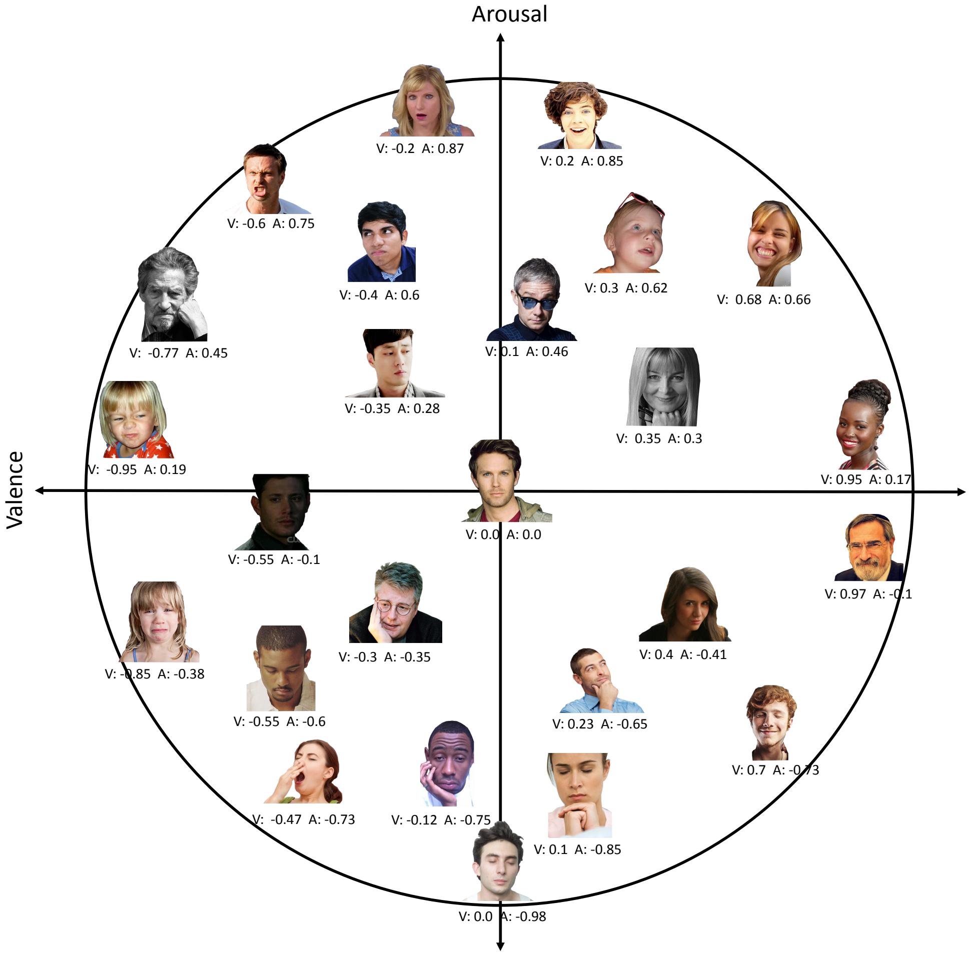


Figure B.1: Sample images in valence arousal circumplex with their corresponding valence and arousal values (V: Valence, A: Arousal).

Table B.4: Number of annotated images in each range/area of valence and arousal

		Valence									
		[-1,-.8]	[-.8,-.6]	[-.6,-.4]	[-.4,-.2]	[-.2,0]	[0,.2]	[.2,.4]	[.4,.6]	[.6,.8]	[.8,1]
Arousal	[.8,1]	0	0	21	674	1021	521	60	57	0	0
	[.6,.8]	0	74	161	561	706	1006	432	738	530	0
	[.4,.6]	638	720	312	505	2689	1905	1228	992	3891	957
	[.2,.4]	6770	9283	3884	2473	5530	2296	3506	1824	2667	1125
	[0,.2]	3331	1286	2971	4854	14083	15300	4104	9998	13842	9884
	[-.2,0]	395	577	5422	3675	9024	23201	6237	42219	23281	21040
	[-.4,-.2]	787	1364	3700	6344	2804	1745	821	5241	10619	9934
	[-.6,-.4]	610	7800	2645	3571	2042	2517	1993	467	1271	921
	[-.8,-.6]	0	3537	8004	4374	5066	3379	4169	944	873	0
	[-1,-.8]	0	0	4123	1759	4836	1845	1672	739	0	0

Table B.5: Evaluation metrics and comparison of CNN baselines, SVM and MS Cognitive on categorical model of affect on the validation set.

	CNN Baselines				SVM	MS Cognitive
	Imbalanced	Down-Sampling	Up-Sampling	Weighted-Loss		
Accuracy	0.40	0.50	0.47	0.58	0.30	0.37
F_1-Score	0.34	0.49	0.44	0.58	0.24	0.33
Kappa	0.32	0.42	0.38	0.51	0.18	0.27
Alpha	0.39	0.42	0.37	0.51	0.13	0.23
AUCPR	0.42	0.48	0.44	0.56	0.30	0.38
AUC	0.74	0.47	0.75	0.82	0.68	0.70

Table B.6: Baselines' performances of predicting valence and arousal on the validation set

	CNN (AlexNet)		SVR	
	Valence	Arousal	Valence	Arousal
RMSE	0.37	0.41	0.55	0.42
CORR	0.66	0.54	0.35	0.31
SAGR	0.74	0.65	0.57	0.68
CCC	0.60	0.34	0.30	0.18

Appendix C

Affect-Aware Agent

C.1 Empathy Stimuli

In order to evoke emotion in subjects, two stimuli each containing four video clips (approximately 30 seconds long) were created. Each clip was intended to elicit a certain spontaneous emotion taken from following YouTube videos:

- Happiness
  - https://www.youtube.com/watch?v=1JArN6rag8s
  - https://www.youtube.com/watch?v=gtMHOot16yk
- Surprise
  - https://www.youtube.com/watch?v=wm0ywsD9V88
  - https://www.youtube.com/watch?v=gtMHOot16yk
- Sadness
  - https://www.youtube.com/watch?v=Gv63u6pFBoI
  - https://www.youtube.com/watch?v=eb3J-GfY6T4
- Disgust
  - https://www.youtube.com/watch?v=QuB3kr3ckYE
  - https://www.youtube.com/watch?v=0sWGpfcQnHk

C.2 Affect-Aware Questionnaire

Table C.1: Affect-Aware Questionnaire

1- The task of describing the videos was easy for me.

1 2 3 4 5
Strongly Disagree Strongly Agree

2- The videos were meaningless.*

1 2 3 4 5
Strongly Disagree Strongly Agree

3- The videos changed my feeling.

1 2 3 4 5
Strongly Disagree Strongly Agree

4- The robot explained the task clearly.

1 2 3 4 5
Strongly Disagree Strongly Agree

5- The experiment was exciting.

1 2 3 4 5
Strongly Disagree Strongly Agree

6- I enjoyed watching the videos.

1 2 3 4 5
Strongly Disagree Strongly Agree

7- The videos were long enough to be described.

1 2 3 4 5
Strongly Disagree Strongly Agree

8- I felt mentally immersed in the experiment.

1 2 3 4 5
Strongly Disagree Strongly Agree

9- Ryan can have tender and concerned feelings for people less fortunate than herself.

1 2 3 4 5

Strongly Disagree Strongly Agree

10- Sometimes Ryan found it difficult to see things from my point of view.*

1 2 3 4 5

Strongly Disagree Strongly Agree

11- Sometimes Ryan did NOT feel sorry for me when I was having problems or issues.*

1 2 3 4 5

Strongly Disagree Strongly Agree

12- If Ryan would see someone being bothered or hurt, she would probably feel protective towards them.

1 2 3 4 5

Strongly Disagree Strongly Agree

13- Ryan sometimes tried to understand me better by imagining how things look from my perspective.

1 2 3 4 5

Strongly Disagree Strongly Agree

14- Ryan is NOT disturbed when I am upset.*

1 2 3 4 5

Strongly Disagree Strongly Agree

15- If Ryan would see someone being treated unfairly, she would NOT feel much pity for them.*

1 2 3 4 5

Strongly Disagree Strongly Agree

16- Ryan is often quite touched by things that she sees happening.

1 2 3 4 5

Strongly Disagree Strongly Agree

17- I would describe Ryan as a pretty soft-hearted robot.

1 2 3 4 5

Strongly Disagree Strongly Agree

18- If Ryan is upset with someone, she would try to put herself in my shoes for a while to understand the situation.

1 2 3 4 5

Strongly Disagree Strongly Agree

19- Ryan tries to imagine how she would feel if she was in my place.

1 2 3 4 5

Strongly Disagree Strongly Agree

20- The robot was friendly.

1 2 3 4 5

Strongly Disagree Strongly Agree

21- The robot was likeable.

1 2 3 4 5

Strongly Disagree Strongly Agree

22- The robot was warm.

1 2 3 4 5

Strongly Disagree Strongly Agree

23- The robot was approachable.

1 2 3 4 5

Strongly Disagree Strongly Agree

* The scores of questions marked with asterisk are reversed.