1-1-2015

Spontaneous Facial Behavior Computing in Human Machine Interaction with Applications in Autism Treatment

Seyedmohammad Mavadati

University of Denver

Spontaneous Facial Behavior Computing in Human Machine Interaction with Applications in Autism Treatment

A Dissertation

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

of Doctor of Philosophy

by

Seyedmohammad Mavadati

March 2015

Advisor: Dr. Mohammad H. Mahoor

Author: Seyedmohammad Mavadati

Title: Spontaneous Facial Behavior Computing in Human Machine Interaction with Applications in Autism Treatment
Advisor: Dr. Mohammad H. Mahoor
Degree Date: March 2015

Abstract

Digital devices and computing machines such as computers, hand-held devices and robots are becoming an important part of our daily life. To have affect-aware intelligent Human-Machine Interaction (HMI) systems, scientists and engineers have aimed to design interfaces which can emulate face-to-face communication. Such HMI systems are capable of detecting and responding upon users' emotions and affective states. One of the main challenges for producing such intelligent system is to design a machine, which can automatically compute spontaneous behaviors of humans in real-life settings. Since humans' facial behaviors contain important non-verbal cues, this dissertation studies facial actions and behaviors in HMI systems. The main two objectives of this dissertation are: 1- capturing, annotating and computing spontaneous facial expressions in a Human-Computer Interaction (HCI) system and releasing a database that allows researchers to study the dynamics of facial muscle movements in both posed and spontaneous data. 2- developing and deploying a robot-based intervention protocol for autism therapeutic applications and modeling facial behaviors of children with high-functioning autism in a real-world Human-Robot Interaction (HRI) system.

Because of the lack of data for analyzing the dynamics of spontaneous facial expressions, my colleagues and I introduced and released a novel database called *"Denver Intensity of Spontaneous Facial Actions (DISFA)"* . DISFA describes facial expressions using Facial Action Coding System (FACS) – a gold standard technique which annotates facial muscle movements in terms of a set of defined Action Units (AUs). This dissertation also introduces an automated system for recognizing DISFA's facial expressions and dynamics of AUs in a single image or sequence of facial images. Results illustrate that our automated system is capable of computing AU dynamics with high accuracy (overall reliability $ICC = 0.77$). In addition, this dissertation investigates and computes the dynamics and temporal patterns of both spontaneous and posed facial actions, which can be used to automatically infer the meaning of facial expressions.

Another objective of this dissertation is to analyze and compute facial behaviors (i.e. eye gaze and head orientation) of individuals in real-world HRI system. Due to the fact that children with Autism Spectrum Disorder (ASD) show interest toward technology, we designed and conducted a set of robot-based games to study and foster the socio-behavioral responses of children diagnosed with high-functioning ASD. Computing the gaze direction and head orientation patterns illustrate how individuals with ASD regulate their facial behaviors differently (compared to typically developing children) when interacting with a robot. In addition, studying the behavioral responses of participants during different phases of this study (i.e. baseline, intervention and follow-up) reveals that overall, a robot-based therapy setting can be a viable approach for helping individuals with autism.

Acknowledgements

First, I would like to acknowledge my incredible advisor, Dr. Mohammad H. Mahoor, who has always been full of inspiration, support and attention. During my PhD studies he always was a great mentor for me and he kindly provided all the required resources and supports to provide a great environment for conducting my research.

I would like to thank my committee members: Dr. Bruce Pennington, Dr. Kimon Valavanis, Dr. Matthew Rutherford, and Dr. Kateri McRae. I feel fortunate to have a diverse committee from Psychology and Engineering backgrounds, who pushed me hard because they believed in me and wanted me to succeed as a scholar and scientist. Moreover, I would like to acknowledge my collaborator Dr. Jeffrey F. Cohn and Mr. Jeffrey M Girard from *University of Pittsburgh*, Dr. Qiang Ji and Dr. Yongqiang Li from *Rensselaer Polytechnic Institute (RPI)* and Dr. Anibal Gutierrez from *Florida International University*. With their help and comments, I could design appropriate protocols and capturing setting for human-machine interaction.

I have had a great fortune to work in a *Computer Vision lab* at University of Denver with a number of talented and smart colleagues and lab mates, who helped me to be a better person and being able to conduct my research in a friendly environment. I am very thankful to all participants that came to our video recording sessions for posed and spontaneous facial expressions study. In addition, I am honored to have the opportunity to interact with children with autism and their parents.

Last but not least, I would love to acknowledge my dad (Mr. Nematollah Mavadati), my mom (Mrs. Zohreh Tababatabaian Nimavard), my older sister (Dr. Paliza) and her husband (Dr. Hadi) and my younger sister (Dr. Shiva) and her husband (Dr. Saeed) and my grand mom (Mrs. Maliheh). I am very fortunate to have such a great family, especially the kind supportive parents who always helped me in every aspect of my life. I would like to dedicate this dissertation to my parents, family members and all of my friends.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the past decades technology has undoubtedly made significant influences on the world and global communications. The emergence of computers and intelligent devices has transformed the society's landscape. In addition, due to the impact of smart computing machines and social networks, the way humans are socializing and interacting with each other has been significantly changed.

Nowadays digital devices are an inseparable part of humans communication. In developed countries, typically every individual regardless of age, gender, and ethnicity interact with machines on a daily basis. The technology ranges from conventional on-line services (e.g. hotel reservations and social networking services) to more advanced systems like on-line tutoring systems and robot-based interfaces. Humans also utilize different types of smart computing devices and technologies (such as cell-phone, tablets and laptops) for various purposes. Due to the nature of human-machine interactions, a new field of research has emerged which is referred to as Human-Machine Interaction (HMI). This area involves the study, design, and application of these interfaces between people (user) and machines. It is worth mentioning, two of the well-growing fields in HMI systems are Human-Computer Interaction (HCI) and Human-Robot Interaction (HRI).

Human-Centered Computing (HCC) is an emerging field in HMI system design, which utilizes the human's facial behavioral cues into the machine control system. For instance, imagine having an intelligent device, which can automatically utilize various forms of verbal and non-verbal social cues (e.g. tone of voice, facial expressions, and body gestures). This capability will help the system better understand or convey the user's messages (See Figure 1.1). At this date, existing technologies do not provide a human-like interaction setting. However, scientists have aimed to move toward this direction by using affect-aware systems.

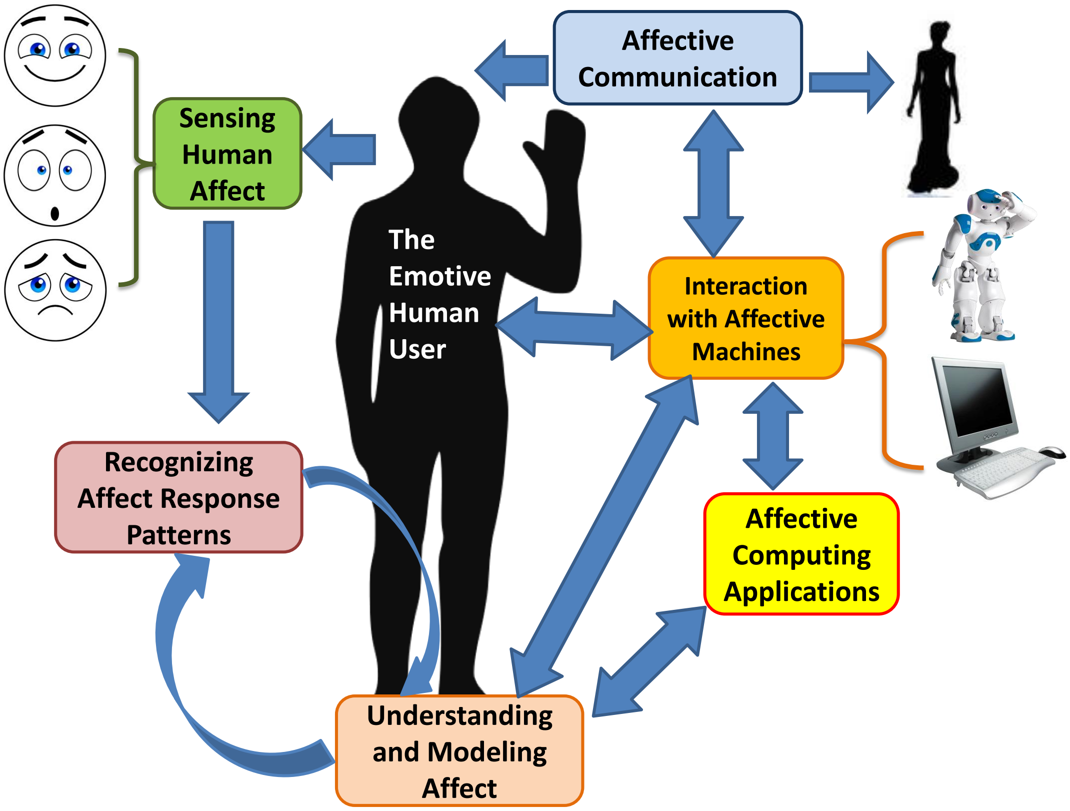

Affective Communication
Sensing Human Affect
The Emotive Human User
Interaction with Affective Machines
Recognizing Affect Response Patterns
Affective Computing Applications
Understanding and Modeling Affect

Figure 1.1: Affect-aware HMI and human-centered computing system [1]

This dissertation focuses on studying facial behaviors of humans interacting with a machine (i.e. a computer or a robot). When I started my PhD studies, the lack of spontaneous real-life data inspired me to capture and annotate related data. In addition, my great passions and interests in understanding machine learning algorithm encouraged me to detect, analyze and compute different forms of affective states. Embedding affective social cue to HMI system design can significantly improve the usability and effectiveness of machines and provide a more comforting and friendly environment for user to interact with machines. The following section introduces the human-machine interaction concepts and provides some background for automated facial image analysis. The contributions and novelties of this study are also stated in the following sections.

1.1 Background

HMI involves studying, designing, and implementing an interface that allows the machine to interact with humans (users). HMI designers aim to improve the experience of users during such interaction and make computers smarter, more usable and receptive toward the user's needs. One of the major goals of HMI designers is to design interactive media that minimizes the obstacles between human cognition and the computer understanding [82].

HMI is the intersection of several different fields of study, such as computer science, engineering, and psychology. Due to the multidisciplinary nature of human machine interaction, researchers with diverse backgrounds are closely working together to have a well-designed, reliable, and easy-to-use platform [82]. One important aspect of designing HMI systems, is the space where interaction between humans and machines occur, which is called *User Interface (UI)*. In order to have a well-designed UI two factors need to be considered: (1) usability and (2) user experience. *Usability* includes the elegance, ease-of-use and clarity of the designed interface. *User experience*, focuses on the user's behaviors, attitudes and emotions, while using the HMI system [82].

Considering the context and application of the HMI, different types of user interfaces may be exploited [155]. Originally for the human machine interaction only peripheral devices and hardware were applied. This category of HMI uses conventional interface devices, such as mouse, keyboard, and LCD monitor and assumes that the user will be explicit, unambiguous and fully attentive during the command flow and information controlling process. This type of interface design functions effectively for context independent tasks, like autonomous reservation system and buying and selling stocks. However, due to the nature of a smart environment and the need for machines to intelligently interact with humans, a new era of interfacing has arisen.

In the new generation of UIs, machines are no longer interconnected with the user through a fixed console. Instead machines use the humans' emotions and responses to adapt their actions appropriately. This category of interface design is more human-centered and aims to improve the usability and user's experience of the system by reading the humans biometrics and signals in a more efficient way [124, 130]. Therefore, the affect-aware UI reads humans commands through linguistic and non-linguistic messages like speech, hand gesture, facial expressions, etc. This type of HMI design is mostly focused on *human computing*, or *human sensing* and provides a natural way of interacting with machine [124]. This version of human machine interaction will preferably be more similar to realistic human-human interactions and would ultimately empower machines to be more involved in human social and daily life [124].

One crucial phase for the latter interface, besides reading the users' commands, is to devise a smart systematic way to train, analyze, and model the emotions and behaviors of the users. This field of research is known as *affective computing* [130]. The human-machine interaction approach introduces a new machine-decision making technique, which requires considering the temporal and emotional states of the user. For instance, using energy level, integrity, attention level, and facial expression patterns of a user can help the machine to sense the user's affective states and based on that knowledge, it decides what types of actions need to be manifested [124, 155]. For instance, in autonomous on-line tutoring systems, the machine may utilize the users' behavioral patterns in addition to the measured attention levels and an engagement factor of the user to automatically adjust the difficulty level of the course and appropriately provide feedback if needed.

One of the most important cues for measuring the humans' moods and behaviors is *spontaneous facial behavior*, such as facial expression, eye contact, gaze direction regulation, and head orientation [155]. Facial behavior analyses can be used for autonomous human machine interaction to better understand the users' emotions. Facial expression helps us better comprehend and conduct a conversation and is undoubtedly one of the most important aspects of nonverbal cues that can be used for autonomous HMI systems. However, to the best of my knowledge, the majority of existing automated Facial Expression Recognition (FER) systems are designed to recognize posed facial actions. The main reason is the lack of an available database which provides a comprehensive annotation for dynamics of spontaneous facial expressions.

To address the aforementioned need, and as the first objective of this dissertation, my colleagues at University of Denver (DU) and I captured, annotated and released a database known as DISFA, to comprehensively analyze the spontaneous facial actions and their dynamics (e.g. presence of AUs, intensity of AUs). Besides, I developed an automated system for recognizing the dynamics of facial muscle movement and AU dynamics. This dissertation also introduces a new posed dataset (DISFA+), which allows us to compare the dynamics and temporal patterns of spontaneous facial actions with the posed AUs. Such comparison enables scientists to better understand the similarities and differences of spontaneous and posed expression and more importantly transfer the learned concepts from the posed facial expressions to the spontaneous domain.

Another objective of this dissertation is to analyze and compute spontaneous facial behaviors (e.g. eye gaze and head direction) of children with ASD in robot-based HMI systems. Such investigations allow us to evaluate how the eye gaze direction and head orientations (which represent the emotions and feelings of humans) are differentiable in ASD and Typically Developing (TD) group. In addition, another goal is to find out how effective robots can interact with children with ASD, especially for for social or behavioral therapy. Hence, I designed and executed a series of robot-based pilot studies for children with autism and coded their facial behaviors and their socio-behavioral responses. The designed protocols, which were a series of social games, were using a robot called NAO. The results and lessons learned from these robot-based therapy sessions can be helpful for those who are interested in employing robots for therapeutic purposes.

1.2 Objectives and Contributions

This section lists the achievements and contributions of this PhD research.

1.2.1 Spontaneous Facial Expression Analysis and Computing

As noted before, one of the main limitations of the existing autonomous FER systems is that they mostly model and classify posed (non-genuine) facial expressions. Spontaneous facial expressions that are observed in HMI settings can be significantly different from the posed ones. The spontaneous expressions may coincide with various head postures, uncontrolled illumination conditions, and subtle facial muscle movements. All these factors can reduce the reliability and the affective recognition accuracy of HMI systems. These challenging factors and limited studies on spontaneous facial expressions encouraged me to investigate this avenue in depth. Below I list the outcomes of my research for different categories of spontaneous facial expressions:

- When I started my PhD in September 2010, there was no available database for modeling the dynamics and intensity of spontaneous facial action units. *To meet the need for a publicly available corpora of well-labeled video of spontaneous facial* *expressions, my colleagues and I collected and ground-truthed the Denver Intensity of Spontaneous Facial Action (DISFA) database* [111]. This database contains the intensity annotations of 12 AUs that have been coded into a six-point ordinal scale. DISFA contains the video facial actions of 27 young adults who were video recorded while viewing video clips that were intended to elicit spontaneous emotion expression. Chapter 2 will introduce facial expression coding, and automatic analysis techniques and explains the video capturing and coding settings of DISFA.

- *As a part of my dissertation, I analyzed the static (frame-based) and temporal (sequence-based) patterns of spontaneous facial actions dynamics (i.e. AU intensity-levels).* Analyzing temporal and intensity dynamics of facial expressions are important steps for categorizing psychological and affective states. Using several different types of appearance features(such as LBPH, HOG and localized Gabor) along with Support Vector Machine (SVM) classifier and Hidden Markov Modeling (HMM) allowed me to computer AU intensities in a single image and sequence of frames of DISFA [110, 111]. Section 3.3, and 3.4 introduces the applied facial expression recognition system and discusses the experimental results of the (static and temporal) autonomous facial expression systems.

- *Developing a software, to capture and annotate the posed facial expressions of 9 subjects (a subset of DISFA database), in order to study and model the differences and similarities of posed and spontaneous facial actions.* Previous studies reported some differences between posed and spontaneous for prototypic expressions (e.g. polite (posed) smile vs joyful smiles) and a few upper face actions [155]. However to the best of my knowledge, none of the existing literature has investigated other prototypic expressions or facial actions. Hence, I developed an application which asks a user to imitate 42 facial expressions for multiple trials. For every participant, the intensity of 12 AUs (reported in DISFA) were annotated. The main objective was to comprehensively study and compare the AU temporal patterns, AU intensity dynamics and co-occurrence of AUs in posed and spontaneous settings. Moreover, I aimed to find out how significantly these characteristics are different in posed and spontaneous ones. Section 3.6 explains the software and recording sessions, FACS annotation, experimental settings and achieved results in more detail.

1.2.2 Autism and Human-Robot Interaction

Preliminary studies in autism research demonstrate that in many cases individuals with Autism Spectrum Disorder (ASD) interact more actively and engagingly with robots than humans. The majority of these studies explore the potential use of social robots to improve the involvement and attention levels of children with ASD. All of these substantial efforts from designing robots to developing robot-based games aim to offer a new doorway to ultimately provide a tool for improving autism therapy alternatives. As there are limited investigations for analyzing facial behaviors of ASD individuals interacting with robots and also utilizing robots in social and behavioral treatments, my colleagues and I designed and executed following robot-based case-studies.

- Eye gaze direction and head pose orientation are important non-verbal communication cues in human interaction. Robot-based studies report inspiring results encouraging the individuals with ASD to communicate with the robot; however, few investigate gaze responses of humans interacting with robots. *I examined gaze responses of children with ASD and Typically Developing (TD) children in conversational context with a robot.* In addition, I automatically tracked the head orientation of subjects during robot interaction, which are helpful measures to specify the attention level of each participant. Section 4.4 and 4.5 report the experimental results of the eye gaze analysis and head orientation of participants.

- As there are limited investigations for utilizing robots in social and behavioral treatments, my colleagues and I designed and executed a therapeutic protocol to conduct a longitudinal case-study using a social robot, such as NAO – a humanoid robot with 25 degree of freedom and remote controlling capabilities [2]). *The objective was to investigate how a social robot (e.g. NAO) can be utilized to effectively improve certain social-emotional behaviors of children with ASD in therapeutic settings.* Results of our treatment procedure specifically tailored to social deficits of each child reveal that the majority of the participants can learn non-verbal and verbal communication skills through social and behavioral treatment. More importantly, they are able to transfer these learned skill concepts to human-human follow up sessions successfully. More details about the robot-based games, capturing setting, and results can be found in Section 4.6.1.

The rest of this dissertation is organized as follows: Chapter 2 introduces the existing facial expression analysis and annotation techniques. Chapter 3 introduces DISFA and DISFA+ databases and presents the experimental results for spontaneous facial expressions analysis, such as static and temporal AU intensity measurement and provides a thorough comparison of spontaneous and posed facial actions. Chapter 4 introduces autism spectrum disorders and the social communication deficits of individuals with ASD. Thereafter, it explains the proposed social robot-based games and the behavioral responses of participant and eye gaze modeling. Chapter 5 concludes the dissertation and outlines future research directions.

Chapter 2

HMI and Automated Facial Behavior Analysis and Computing

Human machine interaction systems have been widely utilized in various applications. The ultimate goal of HMI systems is to develop an easy-to-use product that is effective and enjoyable to users [136]. To design a user-friendly HMI system, we need to know:

- *by whom and where* the HMI system is going to be used?
- *the functionalities and activities* which users are expecting to have while interacting with the system?

For instance to design an efficient communication system, users look for a product which can easily get the input message (i.e. text, voice, visual context) and allows them to comfortably revise with the message (e.g. copy, edit, store, annotate, delete, etc.). In more advanced interfaces, in addition to the easy-to-use platform, it is very important to design an intelligent sensing machine which can compute and analyze human biometrics (e.g. body gesture, humans speech and body language) and can automatically understand humans emotions (i.e. excitement, boredom, tiresome, etc.). Capturing and analyzing these information, empowers the system to be more acceptable and efficient for the users.

Affective computing devices [130] are capable of recognizing human's emotions. Having a machine that have the ability to recognize human's affective states can be useful for variety of applications. For instance the on-line computer-based tutor would be more functional, when it computes user's emotion and changes the machine's responses appropriately. On some applications, such as HRI, robots may show empathy or convey some emotions to users. This make the system more acceptable and friendly. As an example of an affect-aware robot may use different affective channels, such as voice features, tonality, body movement to effectively express its emotions. These HMI interfaces aim to imitate the face-to-face humans communication which makes the HMI systems more acceptable and applicable. HMI system are currently be used in different fields, such as gaming, automated driving vehicles and automatic tennis coach (see Figure 2.1).

Figure 2.1: Human Machine Interaction: Different HMI Interfaces [3, 4, 5, 6, 7]

Humans commonly utilize verbal and non-verbal communication cues for inter-personal communication. Body language is a commonly used non-verbal communication skill, which helps individuals to convey their intentions and physical feelings through their actions and physical behaviors. These behaviors include body gesture, facial expression, head posture, etc. With todays technology, the designers of HMI system aim to design more intelligent systems which can appropriately respond to the users needs. As the facial expression of humans is the main non-verbal channel for socially interacting with others, in this part we will focus on annotating facial expressions and utilizing it for designing an autonomous human affective computing interface.

2.1 Facial Behaviors Analysis: Overview and History

Face is undoubtedly one of the most commonly used means for human-human interaction. We use face to authenticate, and assess others age, beauty, ethnicity and their level of attraction. Scientist [15, 64] considered face as the most important mode of non-verbal communications, which exploits important communication signals, such as facial expression, head movement and gaze direction.

Facial expression analysis has a long history since ancient times, and dates back to the Aristotle and Stewart theories about the facial expressions and functions of emotions. In the late 19th century, Darwin [49] studied the similarity and differences between humans and animals emotions. Duchene utilized electrical stimuli to determine which facial muscle belong to each facial expressions [59]. In the last century, several psychologists and cognitive scientists investigated the area of facial expression analysis in more details.

In early 1970s, Parke (a computer scientist) proposed the first 3D parametric model of the human face to employ a machine to analyze facial images [**?**]. Since that date, there have been several investigations for modeling, representing the facial expressions and designing autonomous systems that can compute expression on videos or still images. This area has gained lots of attentions, popularities and applications in the last two decades. However it is worth mentioning that the majority of studies in this area was limited to investigating and classifying posed facial expression than spontaneous ones. Facial expressions are generally categorized into two groups (See Figure 2.2):

1. **Posed facial expression**, in which a subject is asked to imitate a specific facial action such as joy, anger or other expressions. In this category, the human controls his/her facial muscle movements deliberately, which cause artificial exaggerated expressions.

2. **Spontaneous (non-posed) facial expressions** are those genuinely shown expressions that can be seen in most of people's daily social interactions and individuals do not think when making them.

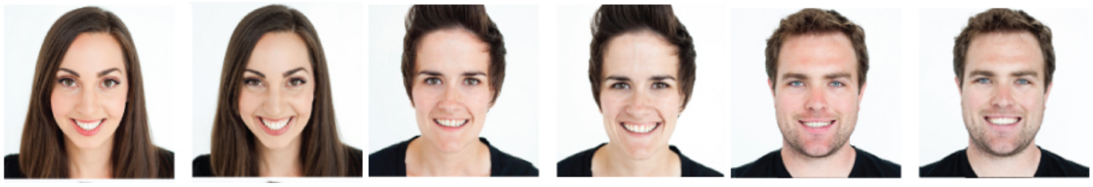


Figure 2.2: Posed (fake) vs. spontaneous (genuine) Smile: Left image in each pair is posed smile and the right one is spontaneous (Duchenne) smile [8].

In order to experience an automated system which can naturally interact with human it is necessary to design an interface which can compute and analyze spontaneous facial expressions. Vast majority of the applications in different fields have encouraged several researchers in the last few years to analyze facial expressions; however it still needs more research and comprehensive investigations. The following sections will review the existing annotation, representation and classification approaches as well as the current facial expression databases.

2.2 Facial Expression Annotation

Paul Ekman (one of the leading and well-known psychologists in facial expression analysis) described two approaches to facial behavior annotations: 1) message-based and 2) sign-based [44]. With *message-based* measurement, the implied meaning of the expression is labeled. Labels for "basic" emotions are frequently used for this category of facial actions [90]; they include *"joy", "surprise", "anger", "fear", "disgust"*, and *"sadness"*, and more recently *"embarrassment"* and *"contempt"* are added to this list (See Figure 2.3). Because humans expressions occur relatively infrequently in their prototypic forms and may mistakenly imply emotions or cognitive-emotional states that are absent [19], more descriptive approaches have been proposed. These are what Ekman refers to as *sign-based* approaches in which the configuration or timing of facial actions is described without reference to the action's presumed meaning [111].

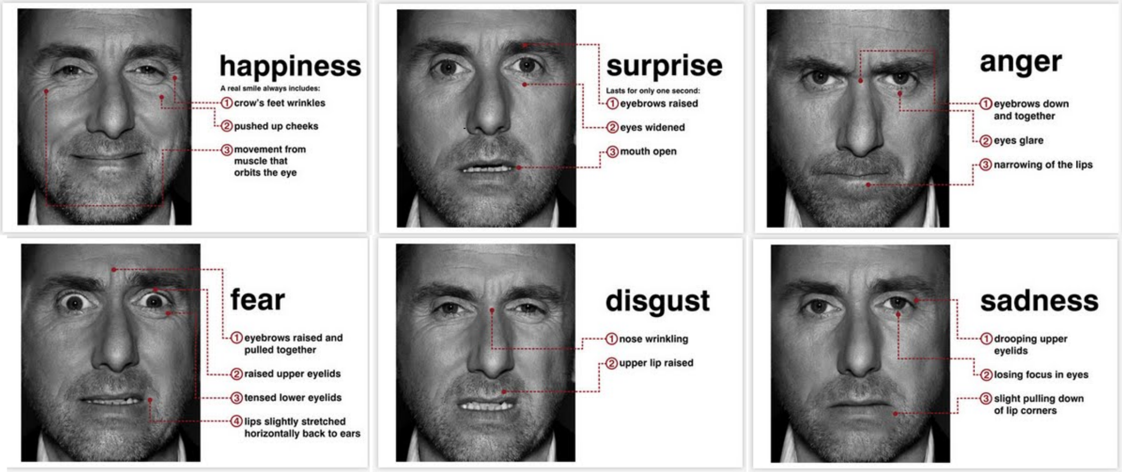


Figure 2.3: Six basic prototypic facial expressions [9]

Since early 1970, facial expression annotation systems have been emerged significantly and several *message-based* and *sign-based* facial behavioral coding techniques have been proposed [139]. For instance, Facial Action Scoring Technique (FAST) [67], A System for Identifying Affect Expression by Holistic Judgment (AFFEX) [86], EMFACS [62], the

Maximally Discriminative Facial Movement Coding System (MAX) [85], and the Facial Action Coding System (FACS) [65] have been introduced and applied; however the chief facial expression annotation techniques are the MAX and FACS.

FACS, informed by the pioneering work of Hjortsjo [81],is more comprehensive than MAX. Facial annotations using MAX [85] provides less complete description of facial actions [109], fails to appropriately discriminate between some anatomically distinct expressions [109], and considers separable expressions that are not anatomically distinct [121]. As the FACS is the most comprehensive existing facial expression coding technique, in the this dissertation I utilize it to describe, annotate, and classify facial expressions.

2.2.1 Facial Action Coding System - FACS

Following seminal efforts by Darwin [49], Duchenne [58] and Hjortsjo [81], Ekman and Friesen [63] developed FACS to describe nearly all possible facial actions. FACS decompose facial expressions into one or more anatomically based FACS Action Units (AUs). No other systems have such descriptive power [68]. In part for this reason, FACS has become the standard for anatomically-based description of facial expression and muscle-based facial animation (e.g. MPEG-4 facial animation parameters [174]).

The 1978 edition of FACS described facial expressions in terms of 44 anatomically based AU and additional action descriptors (ADs) [63]. Action descriptors are actions for which the anatomical basis is either unknown or not specified. A 2002 revision reduced the number of AUs to slightly more than 30, revised the scoring criteria, and improved usability (See Figure 2.4). The coding of intensity was expanded to more AUs and the ordinal scaling of intensity was increased from 3 levels (X, Y, and Z) to 5 (A through E) [66]. Cohn et al. [44] describe the changes made in the 2002 revision of FACS.

FACS employs facial muscle movements and their range of motion to detect the facial actions and their intensities. Every facial expression can be described as a one or a few

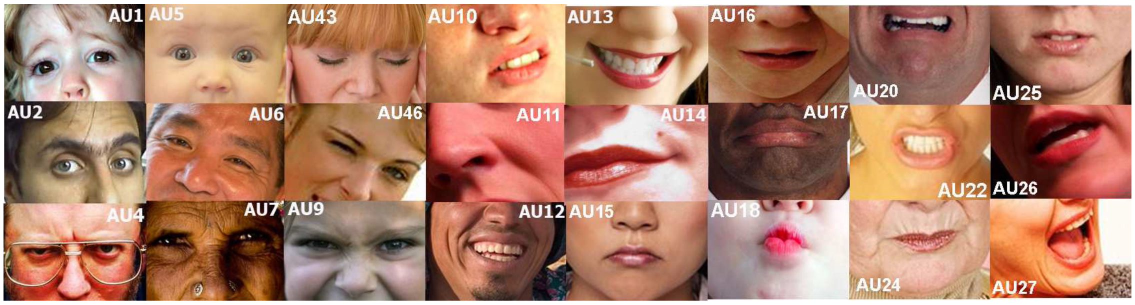


Figure 2.4: FACS action units [63, 155]: **AU1:** Inner Brow Raiser, **AU2:** Outer Brow Raiser, **AU4:** Brow Lowerer, **AU5:** Upper Lid Raiser, **AU6:** Cheek Raiser and Lid Compressor, **AU7:** Lid Tightener, **AU9:** Nose Wrinkler , **AU10:** Upper Lip Raiser, **AU11:** Nasolabial Furrow Deepener, **AU12:** Lip Corner Puller, **AU13:** Sharp Lip Puller, **AU14:** Dimpler, **AU15:** Lip Corner Depressor, **AU16:** Lower Lip Depressor, **AU17:** Chin Raiser, **AU18:** Lip Pucker, **AU20:** Lip Stretcher, **AU22:** Lip Funneler, **AU23:** Lip Tightener, **AU24:** Lip Presser, **AU25:** Lip Part, **AU26:** Jaw Drop, **AU27:** Mouth Stretch, **AU28:** Lip Sucker, **AU43:** Eye Closure, **AU45:** Blink, **AU46:** Wink.

facial muscles movement that may co-occur together. Action units (AUs) provide a terminology for describing nearly all possible facial actions [111]. For example, AU12 (lip corner puller) codes contractions of the zygomatic major muscle and AU25 for the Orbicularis Oris muscle contractions (lips part) [63]. FACS is widely used in behavioral science to investigate emotion, pain, psychopathology, and related topics [68].

By using FACS, expert facial expression coders can determine onset (start or beginning), peak, and offset (stop or end), as well as graded changes in intensity of AUs [66]. FACS assigns the letters A through E to represent intensity variation from barely detectable or trace (A) to maximum (E). As shown in figure 2.5 a given facial expressions can be described as a combination of one or a few AUs with different intensities levels.

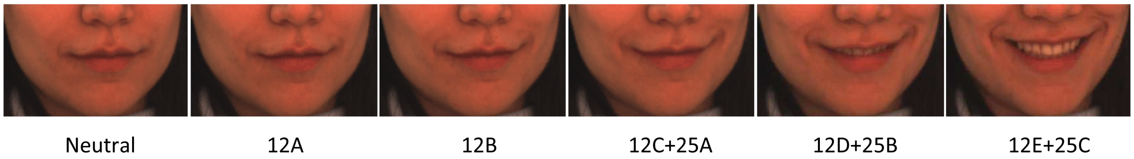


Figure 2.5: Sample facial images with AU intensity variations (AU12: lip corner puller, AU25: lips part) [111].

To annotate facial action units, oftentimes a human expert manually labels every video frame, which is laborious and time consuming. An hour or more may be required to label the onset and offset of each AU for every minute of a video [68]. To label graded changes in intensity, additional time would be required. Although manually annotating facial AUs is standard, it is not ideal, especially when too few manual FACS coders are available for the amount of video to be coded or when instantaneous FACS coding is needed. The ground-truth data is one important phase for designing an automated facial expression computing system which is an essential step toward the real-time HMI applications (like tutoring systems and human robot interaction [57, 164]) that will be described in the following section.

2.3 Automated Facial Expression Analysis

Computing facial expression automatically has remained an active research topic in the last four decades and its focus has been adapted a few times. The very first studies were focused on static images and automatically detecting prototypic expressions (i.e. joy, sadness, surprise) [125, 126, 151]. Thereafter, researchers aimed to investigate and analyze the videos of facial behaviors [126]. In the early 2000s, the new trend for detecting the atomic FACS action units appeared. Recently, the focus has been shifted toward analyzing the spontaneous facial expression and studying the temporal and dynamics of the expressions. In this dissertation, to have a reliable system for HMI interface, the spontaneous facial expression have been studied and modeled.

The majority of existing facial expression studies are focused on the message-based measurements. These investigations have attempted to recognize the affective states (emotions) of human and mostly focus on prototypic facial expressions. Another central points of attention in these studies are related to detecting the cognitive states and energy levels of individuals, such as pain [131, 166], fatigue [76], interest [69] and attention level [83].

The sign-based facial expression (e.g. FACS action unit analysis) investigates the area of research which has been used in a wide range of applications, such as psychology, robotics, health care, education systems, and movie making industries. The main advantage of computing sign-based expressions is that nearly all anatomical and facial muscle movement can be studied independently. In addition, in contrast to message-based analysis the limiting factors such as cultural differences and relation of affective states to the context can also be investigated reliably.

In order to design a reliable facial expression computing system, researchers might aim to design a system which has one or few of the following functionalities:

- being fully automated which can work for both video feeds or images.
- working in real-time and is capable to recognize various spontaneous expressions (i.e. prototypic and facial AUs).
- functioning moderately well in different lighting and occlusion conditions, and being able to recognize expressions from frontal, profile, or other intermediate angels.
- being person independent and invariant to facial hair, glasses, make-up, etc. and can work with different visual resolutions.
- etc.

To design and develop a more intelligent and reliable facial expression computing device, different research groups have been focused on addressing various aspects of the aforementioned points [28]. To meet the need for automated measurement, a number of approaches have been proposed [58, 109, 164]. Examples are modeling the spontaneous and posed expression to distinguish between posed and spontaneous facial expressions [141], and categorizing pain-related facial expressions [17]. We refer readers to [19, 57, 68, 107, 173, 175]. In the following section, we will review the general components for implementing an automated facial expression recognition system.

2.3.1 Components of Automated FER system

Three essential modules for mostly all automated facial expression recognition systems are: 1) Face detection and tracking, 2) Feature Extraction 3) Facial Expression Classification [28].

1) Face Detection and Tracking: Detecting the face in a given image (or a frame of a video) is called ”*Face Detection*” or ”*Face Localization*”, and tracking the detected face across video sequences is ”*Face Tracking*”. Introducing the relevant algorithm for face detection/tracking are beyond the scope of this dissertation. Interested readers can learn more about a few well-known techniques, such as Kanade-Lucas-Tomasi (KLT) tracker [100] and Viola-Jones face detection [158] and Enhanced KLT (EKLT) [30] tracker.

2) Feature Extraction: Feature extraction algorithms mainly convert pixel intensity value of visual data into a higher-level representation (i.e. shape, motion, texture, and spatial configuration of the face or its components). This representation is considered to be a better representation for subsequent expression categorization. Feature extraction algorithms generally introduce high dimensional feature vector, which usually employ the dimensionality reduction techniques to mitigate the ”*curse of dimensionality*” by ideally retaining only essential information with high discrimination power. Statistical, geometric, or spectral-transform-based features are often used as alternative representation of the facial expression prior to classification [37].

3) Facial Expression/ Action Unit Classification: The final stage for facial expression recognition is to apply a classifier (or clustering approach) to discriminate facial features to one of the existing facial actions or expressions. There are several different classifiers, like Naive Bayes (NB), Stochastic Structure Search (SSS) and Dynamic Classifiers (e.g. Hidden Markov Model and Multi-level HMM), Neural network-based classifier and Support Vector Machine (SVM), that can be studied in [28] for more details.

2.4 Literature Review: Automated FER and existing Databases

Beside learning and understanding three deformational components, to design an automated facial expression analyzer we need to have access to a well-labeled database. A suitable database shall contain different facial expressions with various timing and dynamics for multiple subjects. This will help researchers to implement an algorithm and compare their results with state-of-the-art approaches. Section 2.4.1 reviews the existing algorithms for automated FER and Section 2.4.2 reports the facial datasets that are widely used in facial expression community.

2.4.1 Automated Facial Expression Recognition

Automatic computing of facial expression is becoming an important parts of the computer vision and machine learning fields and researchers have become interested in developing algorithms for automatic recognition and analysis of facial expressions [53, 173]. Potential applications of these efforts include human-computer interaction, social robots, consumer photography, automated pain detection, marketing, biometrics, and behavioral and neuroscience. Initial work in some of these application domains is emerging [43, 101, 115, 165].

This section of the dissertation provides an overview of the state-of-the-art approaches for computing facial expressions that are presented in chronological order since 2001. For more comprehensive review the reader is refer to [71, 125]. It is worth mentioning that there are several commonly used databases (e.g. Cohen-Kanade (CK), MMI) for automated facial expression analysis that will be introduced in Section 2.4.2.

In 2001, Tian et al. [150] utilized permanent features (e.g. optical flow, Gabor wavelet, and mutli-state models) and Transient features (e.g. Canny edge detection) for classifying poised facial AUs. They employed Artificial Neural Network (ANN) for both classifying upper and lower facial AUs and their system could classify the facial action with 93% on average. Bourel et al. [31], employed local spatio temporal vectors (obtained from EKLT tracker) and *k*-Nearest Neighbor (KNN) classifier for classifying 25 sequences of four prototypic expressions of CK database.

In 2003, Bartlett et al. [20], applied Gabor wavelet features to the Support Vector Machine (SVM) and AdaSVM classifiers which classified static posed facial expressions with 87% accuracy. Michel and Kaliouby [116], used vector of feature displacement (Euclidean distance between neutral and peak expression) and SVM classifier to discriminate between prototypic posed facial expressions in CK database, with 87% and 71% accuracy for person dependent and independent classification rates respectively.

In 2004, Pantic and Rothkrants [127] employed the frontal and profile facial landmark points in MMI database, using the rule based classifier (86% accuracy). Pantic and Patras [122] also utilized 20 fiducial points besides temporal rules to classify 27 AUs in CK and MMI database. In 2006, they also employed mid-level parameters (using particl filter to track 15 facial landmark points) and rule based classifiers to classify 27 AUs in MMI database [123].

In 2007, Wang and Yin [160] used Topographic Context (TC) expression context descriptor and employed different classifiers (such as Quadratic Discriminant Classifier (QDC), Linear Discriminant Classifier (LDA), Support Vector Classifier (SVC)) to recognize CK and MMI static data. In 2008, Kotsia et al. [93] utilized three types of features (i.e. Gabor Features, DNMF algorithm and geometric displacement vector) to classify six prototypic posed facial expressions, using Multiscale SVM and Multi Layer Perception (MLP).

Besides the posed facial expression analysis, in the last decade several studies have focused on analyzing spontaneous behaviors and facial actions [21, 22, 104]. Analyzing the spontaneous facial expression is a challenging task and currently there are very few studies available in this area. Bartlett et al. [21] measured the intensity of action units in posed and spontaneous facial expressions, by using Gabor wavelet and support vector machines. They have reported the average correlation values of 0.3 and 0.63 between a human coder and the predicted intensity of AUs for spontaneous and posed expressions, respectively. The reported results demonstrate that measuring the intensity of spontaneous expressions is more challenging than measuring the intensity of posed expressions.

Detecting the occurrence of a single or combination of facial action units in static facial images is the area which gain lot of attention [107, 110, 111, 164]. The main focus of the these studies was to utilize investigate the AU intensity of spontaneous facial expressions in static images. [106] presented a framework to automatically measure the intensity of naturalistic AU6 and AU12 of infants using SVM classifier.

Among these studies in automated facial behavior analyses, there are very few literature that has been investigated the temporal facial expressions and measured dynamics of spontaneous facial expressions [96, 137, 147]. [96] utilized a DBN model and SVM classifier to learn the statistical relations among AUs that helps to better classify the AU intensity. [137] employed Markov Random Field (MRF) to estimate the intensity of spontaneous upper face AUs in DISFA database. As mentioned before, one of the challenges for designing spontaneous facial expression recognition system is the lack of well-labeled database. In the next sections, I will introduce the existing and well-known facial expression databases in more details 2.4.2.

2.4.2 Databases for Facial Expression Analysis

To evaluate the efficacy of an automated facial expression analyzer that can detect AUs and may also recognize the dynamics of *'spontaneous'* facial expressions, we need to have a suitable set of data. This section reviews the existing databases which illustrates the lack of publicly available database for studying the spontaneous facial expressions.

In the last two decades, several databases have been designed and released (publicly or privately) for the facial expression analysis. Japanese Female Facial Expression (JAFFE) is one of the firstly released posed datasets, that includes 213 images of 10 Japanese female models expressing seven facial expressions (six basic facial expressions and one neutral) [105]. The Bosphorus data set contains 2D and 3D images of 105 subjects which are coded based on FACS description and are used for posed facial expression recognition and facial action unit detection applications [138]. In 2010, the NVIE database was introduced. This database contains the prototypic facial expressions of 100 subjects in visible and thermal infrared spectrum and each subject participated in three spontaneous experiments and one posed experiment. [161].

For automatic action unit detection, a number of FACS-coded databases have become available for research purposes. Two of the most widely used are Cohn-Kanade [89] and MMI [128]. Cohn-Kanade contains images of 100 subjects that performed 23 facial displays. The image sequences consist of posed facial expressions that include single AU and combinations of AUs in which peak intensity is labeled for each AU. A recent extension of Cohn-Kanade [102] adds additional subjects and spontaneous facial actions. MMI contains over 1500 samples of both static images and image sequences (in frontal and profile views). MMI includes labels for onset (start frame), apex (maximum intensity), and offset (end frame) of each AU but not intensity labels for each frame [128]. MMI also includes some spontaneous facial expressions.

Among other FACS-coded databases are RU-FACS (aka M3) [74] and the UNBC Pain Archive [103]. RU-FACS, which has restricted availability, contains video of spontaneous behavior of 100 subjects that were observed during 2.5 minute interviews [74]. Bosphorus dataset contains 2D and 3D images of 105 subjects and contains both posed facial expression and action units [14]. In [70], Fasel and Luettin utilized a posed database of 9 single AUs and 7 combinations of AUs. The database was FACS coded by Kaiser and Wehrle [88].

The UNBC-McMaster Shoulder Pain Expression Archive [103] provides facial expressions associated with pain. This database includes both FACS and pain intensity coding. Table 2.1 summarizes FACS-coded facial expression databases. Also included are ones that use holistic descriptors such as emotion-specified expressions and valence. Reviewing this table demonstrate that the research community lacks a common ground for computing the dynamics of of spontaneous AUs. This encouraged us to capture, annotate and release a new databased called DISFA [10]. Section 3.1 will introduce DISFA and will explain about the contents and related experimental results on DISFA.

Table 2.1: Description of facial expression databases.

Database	Database information	Expression description	Posed vs Spontaneous	Release(Yr)
Cohn-Kanade [89]	- 100 Subjects (multi-ethnicity) - 69% female, 31% male (18-50 yrs) - Frontal and 30 degree imaging	- AU-coded face database - 23 series of facial display - Single and combinations of AUs - Available for *non-commercial use*	- Posed	**2000**
MMI [128]	- 25 subjects (multi-ethnicity) - 12 female , 13 male (age 20-32 yrs)	- Single and combinations of AUs - Temporal analysis (e.g. onset,apex,offset) - Available for *scientific applications*	- Posed & Spontaneous - Spontaneous, six basic - Expressions added later	**2005**
BU-3DFE [172]	- 100 subjects (multi-ethnicity) - 56 female , 44 male (18 to 70 yrs) - 2,500 3D facial expression models	- Neutral+ 6 basic expressions (4 intensity) - Available for *Research applications*	- Posed	**2006**
JAFFE [105]	- 10 female Japanese models - Grayscale images	- Neutral+6 Basic expressions - 2 to 4 samples per expression - Available for *non-commercial research*	- Posed	**2007**
Bosphorus [14]	- 105 subjects - 44 female , 61 male	- AU-coded (2D/3D data) - Pose and illumination variations - Available for *scientific applications*	- Posed	**2008**
NVIE [161]	- 215 Student (age 17-31 yrs) - Visible and infrared imaging	- Basic facial expressions - Temporal analysis for posed data - Available for *scientific applications*	- Posed &Spontaneous	**2010**
Extended Cohn-Kanade(CK+) [102]	- Extension of Cohn-Kanade - 123 Subjects (multi-ethnicity)	- Onset to peak coded - Spontaneous smiles (66 subjects)	- Posed - Spontaneous	**2010**
UNBC-McMaster Pain Archive [103]	- 129 Participants (63 male,66 female) - 200 video sequence	- Pain related AUs coded - Available for *non-commercial use*	- Spontaneous	**2011**
Belfast Induced Emotion [148]	- Lab-based emotion induction tasks - Three sets of tasks	- Emotion and Intensity (self-report) - Intensity and Valence (trace-style rating)	- Natural emotion	**2012**
RU-FACS [74]	- 100 subjects	- AU-FACS coded (33 AUs) - *Private database*	- Posed & Spontaneous	**2012**
DISFA [10]	- 27 Participants (15 male,12 female) - 130,000 video frames	- Intensity of 12 AUs coded - Available for *non-commercial use*	- Spontaneous	**2013**

2.5 Conclusion

In this chapter I reviewed the existing facial expression annotation approaches techniques (e.g. FACS). Then I introduced the main components of automatic facial expression recognition system and reviewed some of the well-known techniques in this category. I also presented the well-know and mostly used existing databases for facial expression analysis. Reviewing these datasets demonstrate the community needs a new publicly available database with a ground-truthed data for analyzing dynamics of spontaneous AUs. In the next chapter I will introduce our new databases (DISFA and DISFA+) for analyzing AU dynamics and characteristics in both spontaneous and posed setting. The experimental setting and results for computing and modeling the spontaneous facial expressions will be reported. I explicitly report the difference and similarities of facial expression characteristics (temporal and dynamic patterns, AU co-occurrence) for both spontaneous and posed facial expressions and present the results of between-class (posed vs spontaneous) facial expression analysis.

Chapter 3

Automated Facial Expression Analysis

Existing databases and automated system for recognizing prototypic and AU-based facial expressions surveyed in Chapter 2. Reviewing the literature indicates a great need for having a FACS annotated database for spontaneous facial expressions. Therefor, this chapter introduces our recently released database called *Denver Intensity of Spontaneous Facial Actions (DISFA)* which contains the videos and AU annotations of spontaneous facial expressions for 27 subjects. To comprehensively analyze spontaneous facial actions and their dynamics (e.g. presence/absence of AUs, intensity of AUs) I designed and conducted several experimental settings using DISFA database. In addition, this chapter presents a recently captured and annotated database, called DISFA+, which allows us to compare the dynamics and temporal patterns of spontaneous facial actions with posed AUs. Such comparison enables us to better understand and model the similarities and differences between spontaneous and posed expressions and also transfer the learned concepts between the posed facial expressions and the spontaneous domains.

3.1 Denver Intensity of Spontaneous Facial Actions (DISFA) Database

3.1.1 DISFA Contents

This section describes the acquisition setup including eliciting and capturing spontaneous facial behavior for DISFA, FACS annotation and metadata (e.g. facial landmark).

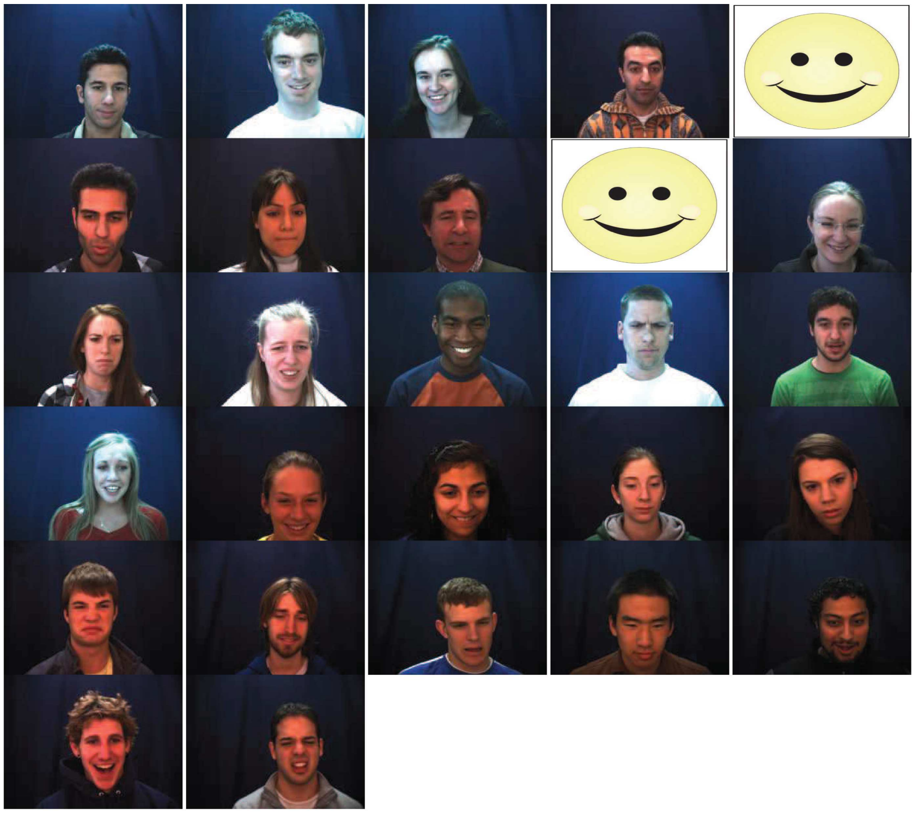


Figure 3.1: Facial images of 25 of 27 participants. Two participants did not consent to use of their images in publications.

Participants

Participants were 27 adults (12 women and 15 men) that vary in age from 18 to 50 years. Three were Asian, 21 Euro-American, two Hispanic, and one African-American.

All participants gave informed consent for the distribution and use of their video images for non-commercial research. Twenty-five of the 27 gave permission for use of their images in publications (Figure 3.1).

Video Stimulus and Recording Procedure

Participants viewed a 4-minute video clip (242 seconds in length) intended to elicit spontaneous AUs in response to videos intended to elicit a range of facial expressions of emotion. In pilot testing, we identified 9 YouTube videos that elicited AUs related to one or more emotions and facial action units of interest. Facial actions of interest were those common in emotion and social communication and that have been a focus of on-going research.

Mindful of time constraints, we selected the most potent segment from each video for inclusion. The 9 segments were concatenated to create the video stimulus clip, which included a one-second pause before each segment. Table 3.1 presents the URL for each clip, its begin and end time, the target emotions, and the percentage of time that each emotion have been occurred. Table 3.2 specifies the emotion description based on [65].

Table 3.1: Description of 9 segments of DISFA stimulus.

Seg. #	**Link**	**Start**	**Stop**	**Target Emotion**	**Emotion Occurrence**
Seg. 1	www.youtube.com/watch?v=qXo3NFqkaRM	00:09	00:20	- Happy	67.3%
Seg. 2	www.youtube.com/watch?v=wwAbtizFCzo	01:00	01:22	- Happy	52.8 %
Seg. 3	· · · /watch?v=1eH9YeZjGIQ&feature=youtu.be	00:19	00:35	- Surprise	9.2%
Seg. 4	www.youtube.com/watch?v=AYsq4fqbMG8	00:48	00:55	- Fear	6.3%
Seg. 5	www.youtube.com/watch?v=QuB3kr3ckYE	00:18 02:09	00:21 03:00	- Disgust	23.6%
Seg. 6	· · · /watch?v=2NEzyvuzs1I&feature=youtu.be	00:08	00:20	- Disgust	5.5%
Seg. 7	· · · /watch?v=DtiSRaB9Sk8&feature=related	01:25	01:59	- Sadness	5.1%
Seg. 8	· · · /watch?v=6PLS7HXo8Bc&feature=plcp	00:00	00:33	- Sadness	1.2%
Seg. 9	www.youtube.com/watch?v=rW29Ex1k2Yc	00:00 01:08	00:15 1:34	- Surprise	11.7%

Table 3.2: Emotion expression in terms of facial action units [65].

Emotion	Criteria
Joy	- Must present AU 12 or AU 6+12 or AU 7+12
Surprise	- Must present AU 1+2+5(low) or AU 1+2+26. (low: Intensity level A or B)
Disgust	- Must present AU 9 or only AU 10
Sad	- Must present AU 1 or AU 1+4
Fear	- Must present AU 1+2+4 or AU 20

While viewing the video, participants sat in a comfortable chair positioned in front of a video display and stereo cameras. They were alone with no one else present. Their facial behavior was imaged using a high-resolution (1024×768 pixels) BumbleBee Point Grey stereo-vision system at 20 fps under uniform illumination. For each participant, 4845 video frames were recorded. The imaging system is depicted in Figure 3.2.

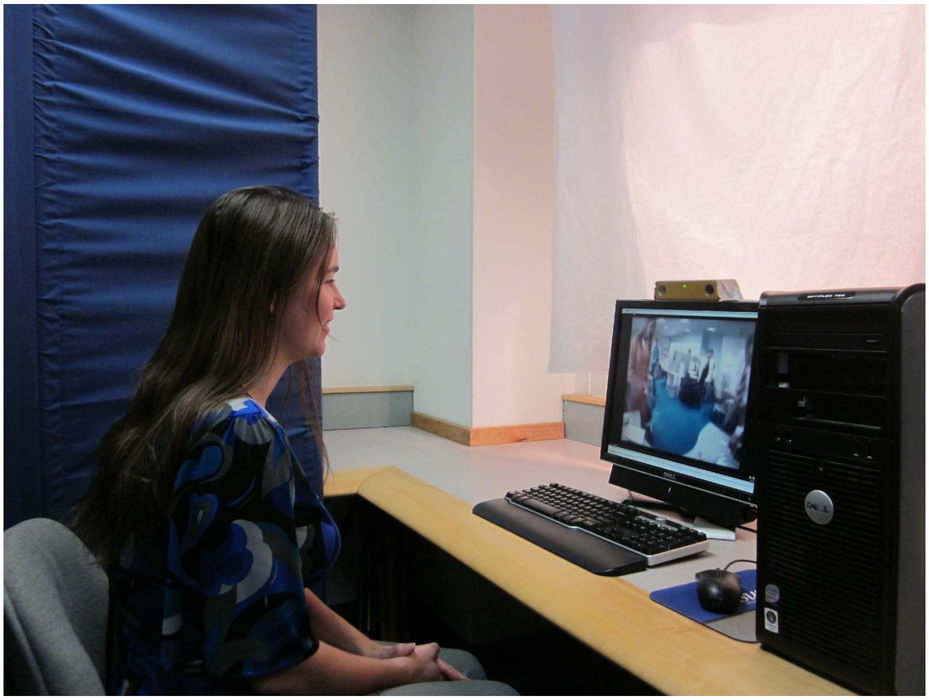


Figure 3.2: DISFA imaging setup.

3.1.2 DISFA Manual FACS Annotation

In contrast to the existing databases, DISFA annotated the FACS intensity levels of 12 spontaneous AUs (i.e. Upper face and lower face AUs). The intensity of each facial action units in DISFA, was coded for each video frame on a 0 (not present) to 5 (maximum intensity) ordinal scale [65]. For each AU, we report the number of events and the number of frames for each intensity level. Event refers to the continuous occurrence of an AU from its onset (start frame) to its offset (end frame) (see Table 3.3 for more details).

Table 3.3: The number of AU events (EV) and total number of frames for each intensity level.

	AU Info		**# Frames for each AU Intensity**					
	AU#	**#EV**	**0**	**1**	**2**	**3**	**4**	**5**
Upper-Face	1	161	122036	2272	1749	2809	1393	555
	2	111	123450	1720	934	3505	836	369
	4	238	106220	4661	7636	6586	4328	1383
	5	100	128085	1579	719	293	104	34
	6	170	111330	9157	5986	3599	601	141
Lower-Face	9	73	123682	1659	2035	3045	316	77
	12	250	100020	13943	6869	7233	2577	172
	15	97	122952	5180	1618	1017	47	0
	17	271	117884	6342	4184	2281	112	11
	20	99	126282	1591	1608	1305	28	0
	25	296	84762	9805	13935	15693	5580	1039
	26	321	105838	13443	7473	3529	314	217

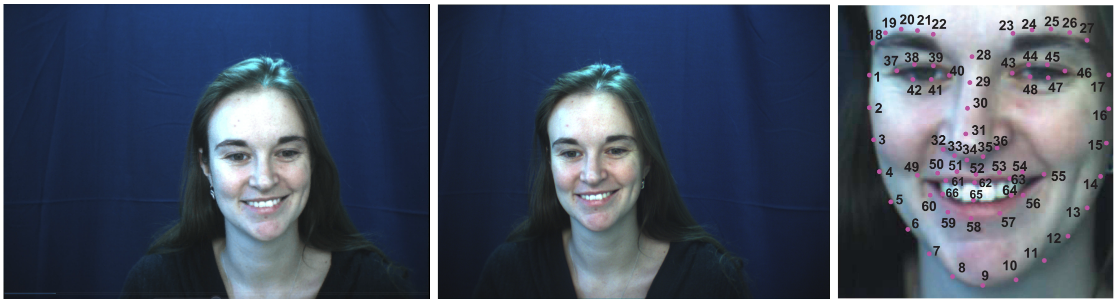
In order to code the facial expressions of every frame of video in DISFA, the FACS coding was performed by a single FACS coder. To evaluate the performance of manual coding, video from 10 randomly selected participants was annotated by a second FACS coder and the inter-observer reliability for all AUs were calculated. Both coders were certified in use of FACS. Inter-observer reliability, as quantified by intra-class correlation coefficient (ICC), ranged from 0.80 to 0.94 (as seen in Table 3.4). ICC value of 0.80 and higher is considered as high reliability [42]. The results demonstrates the high reliability of the AU annotations in DISFA.

Table 3.4: AU description and inter-observer reliability.

AU	Description	Reliability (ICC)
1	Inner Brow Raiser	0.83
2	Outer Brow Raiser	0.86
4	Brow Lowerer	0.94
5	Upper Lid Raiser	0.81
6	Cheek Raiser	0.80
9	Nose Wrinkler	0.85
12	Lip Corner Puller	0.90
15	Lip Corner Depressor	0.82
17	Chin Raiser	0.81
20	Lip Stretcher	0.83
25	Lips Part	0.92
26	Jaw Drop	0.93

3.1.3 AAM Facial Landmark Points

Most techniques for automatic facial expressions recognition and AU detection require a set of landmark points marked in facial images. These points may be used for facial image registration, head pose estimation and compensation, and high-level feature extraction. To detect and extract landmark points, we used active appearance models (AAM) [46] which is a powerful algorithm for matching a statistical model of appearance and shape of an object to a new given image. We extracted 66 landmark points in each image. The (x, y) coordinates of these 66 points have been provided for every frame of DISFA database.



(a) Left Camera

(b) Right Camera

(c) AAM Landmarks

Figure 3.3: (a) Left view, (b) Right view, (c) Sixty-six Landmark points on the face

3.2 Automatic Action Unit Intensity Measurement

As introduced in Section 2.3.1, three essential components for automatic facial expression recognition are: 1) Face detection and tracking, 2) Feature Extraction 3) Facial Expression (AUs) classification. These three steps are explained bellow in more details.

3.2.1 Facial Image Registration and Normalization

To automatically recognize the facial expressions and different AUs in DISFA, we employed the 66 landmark points around facial components (see Section 3.1.3). These landmarks are utilized to register every frame of DISFA to the predefined frontal model. We employed a similarity transformation to map facial landmark points to a canonical orientation. Using the transformed landmarks, a binary mask then was fitted to the image to remove non-face background (see Figure 3.5). The cropped facial regions were resized into 108×128 pixels and in the our experiments we used this preprocessed images for feature extraction and classification.

3.2.2 Feature Extraction

For our experiments, three sets of appearance features were extracted from the cropped and aligned images. These are local binary pattern histogram, histogram of oriented gradient, and localized Gabor features, as described below.

Local Binary Pattern Histogram (LBPH)

The Local Binary Pattern (LBP) codifies local texture information of an image. It is a well-known technique for object representation and classification and is widely used for biometrics [13] and facial expression recognition[143]. An important property is that it is robust to variation in illumination [120]. LBP histogram (LBPH) is the histogram of LBP labels computed over a region as a texture descriptor. With LBPH, there is an inherent trade-off between amount of retained shape information and the dimensionality of the data.

To calculate the LBPH, we defined a rectangular region (i.e. 18×16 pixels) around every landmark point and then used a uniform LBPH feature with 59 bins [143]. By concatenating the LBPH features for each region, we obtained a feature vector of length 3894.

Histogram of Oriented Gradients (HOG)

The histogram of oriented gradient (HOG) descriptor counts the occurrences of gradient orientations in a localized portion of an image. This technique was introduced by Dalal and Triggs for the application of human detection [48], and since then has attracted attention in the face analysis field [146]. HOG represents both appearance and shape information, which makes it suitable for representing different facial expressions. To represent spatial information using this technique, images are divided into small cells; for each cell, the histogram of gradients is calculated [48].

We defined a cell of size 18×16 pixels around every landmark point, for a total of 66 cells for each image. We also applied the horizontal gradient filter $[-1 \ 0 \ 1]$ with 59 orientation bins [48]. In order to form the HOG feature vector, the HOG representation for all the cells were stacked together. This resulted in a HOG feature vector of size 3894.

Localized Gabor filters

Gabor filters are another widely-used technique for representing facial textures in face and expression recognition [23, 108]. To extract a global appearance of a face, Gabor filters are applied to the whole face. We extracted local features by applying Gabor filters to specific regions of the face (e.g. around facial landmark points). We applied 40 Gabor filters (5 scales and 8 orientations) around each of 66 landmark points, which resulted in 2640 Gabor features.

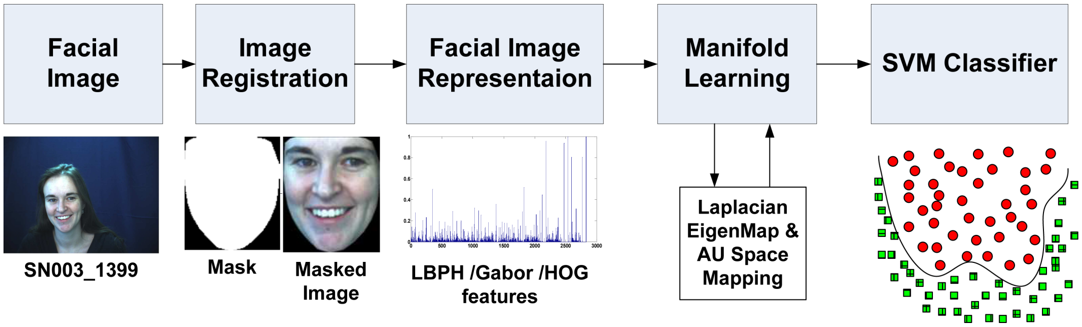

Facial Image
Image Registration
Facial Image Represenation
Manifold Learning
SVM Classifier
SN003_1399
Mask
Masked Image
LBPH /Gabor /HOG features
Laplacian EigenMap & AU Space Mapping

Figure 3.4: Benchmark system for automatic AU intensity measurement.

3.2.3 Manifold Learning for Feature Dimensionality Reduction

In majority of feature extraction techniques, the size of feature vectors is too larger than available size of training samples, which is referred to as *curse of dimensionality*. In order to be able to efficiently train an accurate an robust classifier, we employ a feature dimensionality reduction technique, which can be classified into 1) Linear (e.g. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA),etc) 2) Non-Linear (e.g. Kernel PCA, Manifold Learning) are utilized.

In our experiments, to reduce the high dimensionality of the features, we used non-linear *manifold learning*. This approach assumes that the features lie in a low-dimensional manifold embedded into a high dimensional space. Mathematically speaking, given a set of points $x_1, ..., x_n \in R^D$ find a set of points $y_1, ..., y_n \in R^d$ ($d << D$), such that y_i represents x_i efficiently [35]. There are several different manifold learning techniques that has been introduced since early 2000, and we employed 'Laplacian Eigenmap'. For more details we refer readers to [27, 35].

The *Laplacian Eigenmap* [27] followed by spectral regression (SR) technique was used to extract low-dimensional features of facial images. We randomly selected a small number of images from all subjects to learn a separate manifold for each action unit. Similar to our previous work [106], we applied the spectral regression algorithm to calculate the projection function for each action unit, and then used it to transform high-dimensional LBPH, HOG, and Gabor features into separate low dimension spaces for building a more efficient classifier.

3.2.4 Classification

The support vector machine is a classification approach that has gained popularity for pattern classification in the last decade. Relative to some other approaches (e.g. discriminant analysis and logistic regression), it can be applied to both linear and non-linear classification and is computationally relatively efficient. The SVM classifier applies the kernel trick, which uses dot product, to keep computational loads reasonable. The kernel functions enable the SVM algorithm to fit a hyperplane with a maximum margin into the transformed high dimensional feature space. Several kernel functions have been proposed. Some of the most popular kernels are *Linear* kernel, *polynomial* kernel, *Gaussian radial basis function* (RBF) kernel, and *Sigmoid* kernel. SVM is originally designed for binary classification however it can be applied for multi-class SVM too. To classify C classes of data, two alternative approaches are used. One is *one-against-rest* strategy in which $C - 1$ different classifiers are defined to determine whether a data point belongs to class C_i or to other remaining classes $C_{j \neq i}$. Another strategy is *one-against-one* in which $C(C-1)/2$ binary discriminant functions are defined, one for every possible pair of classes. More details about SVM can be found in [156].

We used a multi-class SVM classifier for AU intensity measurement. We examined three different kernels (i.e. Linear, Polynomial, and Gaussian RBF). The Gaussian RBF kernel outperformed the other two kernels. To train SVM classifiers, the one-against-one strategy was utilized (i.e. 15 classifiers were trained for six-intensity levels of each AU). Test samples were assigned to the class with the highest probability. The schematic framework for AU measurement is shown in Figure 3.4.

3.2.5 Concurrent Validity of the System

Any measurement is associated with some error. Because error can be variously defined, several reliability indices have been proposed. Widely used in automated facial expression detection and behavioral science are F1-score [149], Cohen's-Kappa [41], and intraclass correlation (ICC) [145]. F1 and Kappa are appropriate for nominal data (e.g. presence/absence of AU) and ICC for ordinal or interval-level data (e.g. AU intensity).

ICC is used to calculate the reliability of judgments in multi-class recognition applications. It thus may be used to quantify reliability for AU intensity, which is measured on a 0 to 5 scale. ICC has several definitions that may yield different results for the same data. Selecting the best ICC technique is not possible without knowing the judgment model of the problem. For instance, to model *n* instances by *k* judges, two possible models are: 1) each target is rated by *k* judges, each of which is randomly selected from a larger set of judges, and 2) each target is rated by the same *k* judges of interest. [145] discusses these and other measurement models.

In our case, the second model applies (equivalent to ICC(3, 1) in [145]). There were *k* judges and *n* targets, and by defining the within-target sum square (WSS), between raters sum squares (RSS), between-target sum squares (BSS) and residual sum of squares ($ESS = WSS - RSS$) the ICC can be calculated by:

$$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS} \tag{3.2.1}$$

where $BMS = \frac{BSS}{n-1}$ is between-targets mean square and $EMS = \frac{ESS}{(k-1)\times(n-1)}$ is the residual mean squares [145]. ICC is a powerful reliability measure for multi-class classification problems because it takes into account the difference between large and small errors between judges (e.g. predicting label '5' for a true label '4' has higher ICC than prediction '1' for the true label '4').

3.3 Static Analysis: AU Intensity Measurement

For static analysis we aimed to analyze the facial expression dynamics, using a single image (i.e. one frame of the video). In this study, we detected the intensity of 12 AUs on a 0-5 scale using three facial representations, LBPH, HOG, and Gabor. We used a leave-one-subject-out (LOSO) cross validation technique in which a subset of the videos of 26 subjects was used for SVM training (i.e. about 3000 randomly selected video frames). Testing was performed on the left-out subject. This scenario was repeated for every subject and finally the accuracy and ICC are reported. Accuracy was defined as an average of correctly detected intensity levels for each AU. In our experiment we utilized LIBSVM library with parameters $[-s\,0, -t\,2, -c\,32, -g\,0.1]$ (refer to [36] for definition of LIBSVM parameters).

As shown in Table 3.5, LBPH and HOG features led to accuracy less that 82% and ICC value about 0.70. Overall, the Gabor feature had 86% accuracy and 0.77 ICC value, which yielded the best result among the three facial representations. Table 3.5 shows the accuracy for each of the 6 intensity levels.

Comparing Table 3.3 and Table 3.5 suggests the importance of having sufficient samples for each intensity level. Detection metrics are lowest for those intensity levels that occur least frequently. For instance, AUs, such as AU12 and AU25, have large number of samples for all six levels of intensity, and performance metrics are generally high (e.g. $ICC > 0.84$). AU20, on the other hand, has relatively sparse rates of occurrence for some intensity levels. Performance metrics for AU20 are lower than those for AU12 and AU25. Because all three involve similar amounts of change in appearance, it is likely that these differences are due to the uneven representation of AU intensity for AU20 (i.e. $ICC \approx 0.55$). In the next step we will focus on analyzing the dynamics of spontaneous facial actions and try to classify them using a sequence of visual data.

Table 3.5: AU intensity measurement results; (a)LBPH, (b)HOG, (c)Gabor features.

(a) LBPH features

		AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU15**	**AU17**	**AU20**	**AU25**	**AU26**	**AVG.**
ICC		0.75	0.74	0.83	0.50	0.73	0.65	0.83	0.60	0.53	0.49	0.91	0.72	**0.69**
Accuracy(%)		83.22	85.34	78.31	90.79	78.84	89.67	72.94	88.97	74.04	88.98	72.73	74.67	**81.54**
6-Level Intensity Accuracy (%)	’0’	85.82	87.31	84.33	91.80	83.98	92.16	77.83	91.69	77.40	90.56	78.41	80.45	**85.14**
	’1’	33.08	27.73	39.15	34.78	43.83	18.93	53.95	39.62	42.94	33.02	53.99	46.71	**38.98**
	’2’	58.50	67.03	47.27	54.19	53.14	53.18	63.95	58.40	51.45	70.16	59.28	55.46	**57.67**
	’3’	48.67	55.71	53.45	54.30	65.72	58.83	52.11	68.13	37.07	28.75	64.22	52.77	**53.31**
	’4’	52.26	47.34	70.35	74.75	53.24	63.84	79.42	80.00	1.79	32.14	74.79	44.40	**56.11**
	’5’	49.73	90.81	56.74	91.18	0.00	39.66	12.79	-	0.00	-	87.23	62.79	**49.09**

(b) HOG features

		AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU15**	**AU17**	**AU20**	**AU25**	**AU26**	**AVG.**
ICC		0.75	0.74	0.84	0.58	0.77	0.74	0.75	0.65	0.57	0.46	0.79	0.69	**0.70**
Accuracy(%)		80.24	83.92	75.31	92.85	80.70	91.97	65.55	87.58	72.56	84.91	60.98	73.28	**79.14**
6-Level Intensity Accuracy (%)	’0’	82.23	85.36	79.65	93.78	86.29	93.86	71.91	89.48	75.06	85.99	66.89	77.41	**82.33**
	’1’	34.63	26.25	37.67	39.43	43.99	21.54	40.93	57.98	40.12	43.16	38.80	50.57	**39.59**
	’2’	63.28	42.56	54.49	71.68	45.01	58.62	40.76	64.52	56.72	60.74	50.93	63.31	**56.05**
	’3’	50.96	83.07	56.00	54.64	76.12	73.80	52.72	52.79	69.46	60.88	58.75	58.82	**62.33**
	’4’	59.38	64.52	71.82	47.47	51.75	64.17	65.80	22.22	6.25	46.43	44.08	61.19	**50.42**
	’5’	83.24	36.77	81.81	82.35	0.00	27.59	0.00	-	0.00	-	51.67	73.26	**43.67**

(c) Gabor features

		AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU15**	**AU17**	**AU20**	**AU25**	**AU26**	**AVG.**
ICC		0.80	0.83	0.87	0.58	0.81	0.80	0.84	0.71	0.69	0.54	0.94	0.79	**0.77**
Accuracy(%)		86.24	91.28	80.94	94.46	83.98	92.32	79.67	91.35	80.66	88.12	79.85	79.61	**85.71**
6-Level Intensity Accuracy (%)	’0’	88.06	92.45	84.36	95.49	87.69	93.61	84.55	93.12	82.71	89.02	82.11	82.52	**87.98**
	’1’	52.55	61.83	57.35	36.20	58.16	75.06	63.65	62.81	56.63	74.95	70.11	63.11	**61.03**
	’2’	77.74	66.26	69.10	66.03	65.69	64.62	68.97	59.83	69.93	55.53	76.25	68.03	**67.33**
	’3’	61.92	86.20	70.93	54.64	77.83	74.39	64.09	82.07	65.71	59.57	73.81	79.08	**70.85**
	’4’	64.49	56.98	64.82	53.54	59.57	60.26	59.72	0.00	46.43	7.14	85.97	85.07	**53.65**
	’5’	31.17	28.13	59.58	55.88	0.00	15.52	0.00	-	0.00	-	96.20	93.60	**38.01**

3.4 Temporal Analysis: AU Intensity Recognition

Recognizing the temporal and dynamics (i.e. Onset, apex, offset) of spontaneous facial expressions automatically, has wide range of applications, such as designing an adaptive on-line tutoring platform, developing a friendly human machine interface and customizing a computer-based therapeutic intervention for autistic or schizophrenic individuals [77, 98, 164]. With few exceptions [40, 96, 97, 152], majority of the previous studies in the area of facial expression recognition have focused on recognizing expression in a static image. In other words, the facial behavior videos has been treated as independent set of images. Moreover, the majority of these studies have been analyzing the prototypic facial expressions (happiness, fear, anger, etc.) [40] and posed facial actions [97, 152].

In order to automatically recognize the spontaneous facial expressions in the wild (i.e. un-controlled capturing setting with various illuminations, head pose variations, etc.) it is critical to utilize temporal structure of facial behaviors. In this section we will describe our proposed AU intensity measurement using temporal information, but prior to that some of the related works in this area will be reviewed.

3.4.1 Literature Review: Temporal Facial Expression Modeling

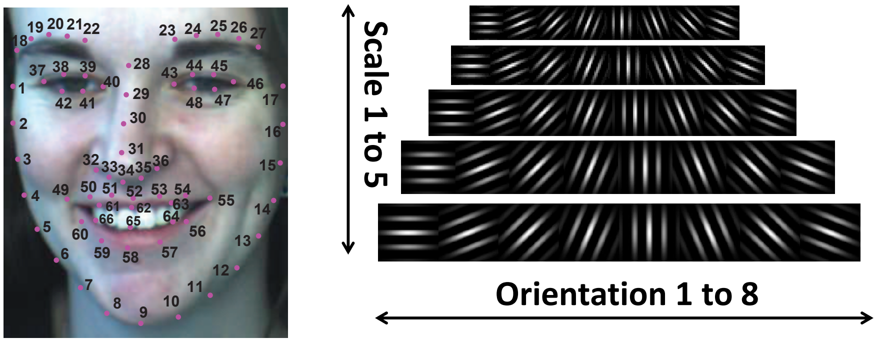
In last decade, a few approaches have been proposed to employ the temporal information of posed facial expression for prototypic facial expression recognition, AU detection and dynamic modeling [52, 152]. One of the pioneering studies in the area of automated AU-based facial expression recognition has been proposed by Lien et al. [97]. They introduced appearance and motion based features from a sequence of images and applied them to Hidden Markov Model (HMM) for recognizing a group of posed AU combinations. [40] employed Baysian network and HMM classifier to learn dependencies among different facial features in videos and designed a system to classify six prototypic posed expressions.

Tong et al. [152] employed Dynamic Bayesian Network (DBN) to model and analyze the dependencies among posed facial behaviors and employed it for AU detection application. In 2010, Schmidt et al. [142] proposed an algorithm to extract appearance, shape and motion features of posed prototypic expressions and apply HMM to classify basic facial expressions .

The main focus of the aforementioned studies was to utilize temporal information of images sequences to classify either posed facial action units or prototypic facial expressions. There are very few literature that have been investigated the AU intensity measurement and dynamics of spontaneous facial expressions. [106] presented a framework to automatically measure the intensity of naturalistic AU6 and AU12 of infants using SVM classifier. As described in previous section, Mavadati et al. [111] employed Gabor feature and manifold learning technique to represent static facial expression and utilized multicalss SVM classifier to measure intensity of twelve spontaneous AUs. In order to address the challenges of spontaneous facial expression analysis (e.g. head pose variations, subject-independent classification, subtle facial expression) we designed an automated system which exploits temporal structure of facial behaviors for recognizing the intensity of spontaneous AUs. The aim of this section is twofold. First we model the temporal patterns of facial actions using statistical Hidden Markov Model (HMM) [87]. In this phase, we quantitatively measure the effects of inter-subject and intra-subject temporal pattern variability on recognizing facial actions. Second we propose to model and classify the intensity of all twelve spontaneous facial actions in DISFA [111] using HMM-based analyzer. In the experiments we also evaluate the effects of timing for recognizing the intensity of different facial actions.

3.4.2 Experiments: Temporal AU Dynamics Modeling

To modeling the AU intensities temporally we utilized DISFA (i.e. 12 AUs, AU intensity in a range of $[0 - 5]$). In our experiments we employed the 66 AAM landmark points along with the similarity transformation to map facial landmark points to a canonical orientation. Moreover we utilized Gabor filters for integrating both shape and texture information in the feature vector. We applied 40 Gabor filters (*5 scales* × *8 orientation*) on every registered facial images (see Figure 3.5(b)). In order to extract discriminative features that can represent the AU intensity effectively we selected a set of landmark point and extracted Gabor features at these points.



(a) AAM landmarks (b) Gabor Filters

Figure 3.5: (a) 66 Landmark points, (b) 40 Gabor Filters which applied to the landmarks

To model the temporal structure of facial expression Hidden Markov Model was utilized (See Appendix A.1). We followed the shown framework in Figure 3.6 and to measure and analyze the intensity of AUs we conducted two experiments:

Experiment 1: Subject Dependent

The goal of this part is to analyze the accuracy and reliability of HMM to model and recognize temporal patterns of inter-subject facial expression. We randomly divided sequence of facial images of every subject into training and testing sets. The training set was used for vector quantization, symbols definition and setting HMM parameter values (One HMM for each AU intensity as explained in Appendix A.1). In our experiments the number of states and length of observation was fixed to $N = 5$ and $T = 10$ respectively.

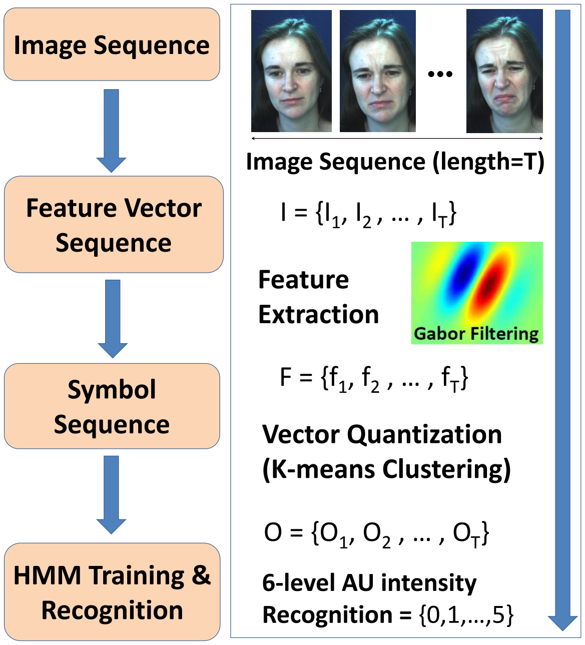


Figure 3.6: AU intensity recognition Using HMM

In the testing phase each frame of video sequence was represented by one symbol, therefore we coded the consecutive frames by sequence of symbols. In order to classify a video segment into one of the AU intensity categories, the likelihood of the symbol sequence with all HMMs are calculated. Then the sequence has been assigned to a class which has the highest likelihood value. The accuracy of AU intensity measurements for all 6 intensity levels and the Intra Class Coefficient (ICC) [111] of all AUs have been reported in Table 3.6. ICC is a descriptive statistics that works on data structure as a group and has a value within range of [0-1]. Higher ICC values illustrates more resemblance between true and predicted AU intensities.

The results in Table 3.6 illustrate that subject-based AU intensity measurement in overall has high accuracy (over 80%), and moderate ICC value about 0.65. Moreover, HMM can accurately recognize some of the AUs with high intensity (e.g. AU1, AU2, AU25 and AU26). This can be claimed as the results of sharp variation in texture and shape variations for illustrating the facial expressions with presence of these AUs. For AU5 it can be seen that the automated system has low reliability ($ICC \approx 0.45$). This may suggests that the performance of our HMM AU intensity analyzer is sensitive to the number of training samples (see number of samples of AU5 for different intensity levels in Table 3.3).

Experiment 2: Subject Independent

In this part of experiment we used two independent sets of subjects for training and testing HMM (akal subject independent experiment). The goal of this study is to investigate which AUs and in what sense AU dynamics have universal temporal pattern (i.e. same patterns among different subjects). In order to have subject independent configuration we employed leave-one-subject-out –LOSO– cross validation configuration. This means that 26 subjects of DISFA were used for designing HMM classifier and use the excluded subject for testing the model. These experiments have been conducted for all 27 subjects and the overall accuracy and ICC values have been reported. Subject independent experiment is more applicable for modeling the temporal patterns of facial expressions for AU intensity analysis in the real world as the training and testing set are totally independent.

The results of subject independent temporal modeling (shown in Figure 3.7 and 3.8) demonstrate that the recognition and ICC rates are not as high as the previous experiment. This means that generally speaking temporal patterns of AU intensities in each person has some uniqueness. This means that the timing for showing different AUs are varying among subjects.

Furthermore to illustrate the effects of length of sequence (T) in HMM modeling in our experiment we reported the recognition results for five values of T (T= [8 10 12 16 20]). The experimental results show that for majority of upper face AUs (e.g. AU1, AU2,AU4,AU5,AU6), HHM has higher accuracy for the sequence with smaller length.

This specifies that upper face AUs have rapid temporal variations than lower face AUs and smaller T can model their temporal variations more accurately.

Table 3.6: Subject dependent results of AU intensity measurement for all 12 AUs of DISFA

		AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26	AVG.
ICC		0.75	0.71	0.76	0.45	0.62	0.53	0.78	0.49	0.58	0.47	0.90	0.67	**0.65**
Accuracy(%)		85.53	88.23	76.13	91.58	78.35	86.51	76.63	87.85	81.35	89.50	76.66	78.17	**83.05**
6-Level Intensity Accuracy (%)	'0'	89.03	91.49	83.31	92.88	83.60	88.49	82.68	89.78	85.59	91.38	82.46	84.31	**87.08**
	'1'	20.79	33.33	43.82	42.85	50.31	48.71	55.37	60.24	48.70	44.61	59.45	53.75	**46.83**
	'2'	29.85	19.56	51.40	33.33	56.94	52.00	54.79	49.23	50.66	39.65	66.04	56.72	**46.68**
	'3'	59.25	53.06	36.31	50.00	47.86	69.79	63.56	75.00	60.97	58.49	68.76	61.90	**58.76**
	'4'	65.62	51.61	51.38	100.00	48.27	50.00	70.65	100.00	16.67	0.00	76.23	36.36	**55.57**
	'5'	94.73	85.71	64.81	0.00	25.00	33.33	66.66	-	0.00	-	83.33	71.42	**52.50**

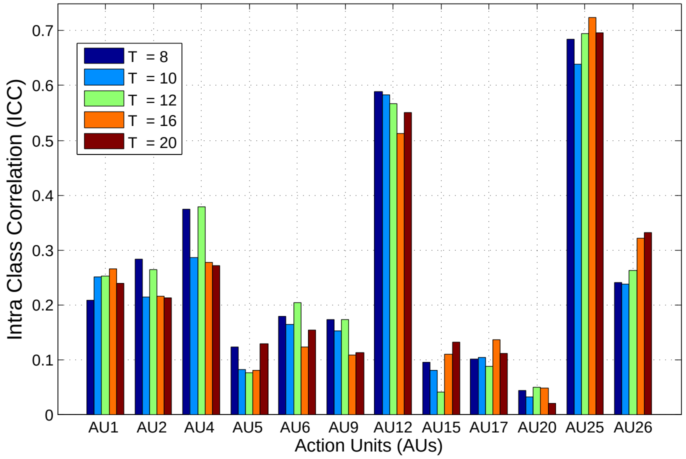


Figure 3.7: Overall ICC of subject independent experiment: effects of sequence length (T) on AU intensity reliability

.

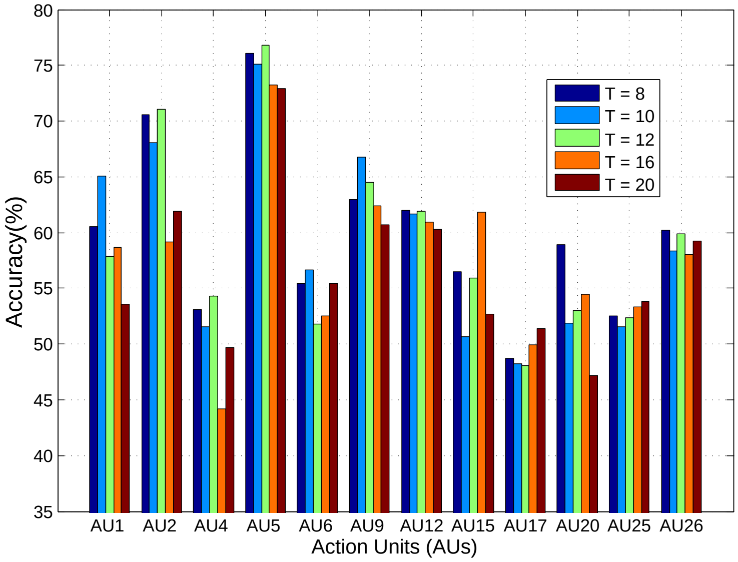


Figure 3.8: Overall accuracy of subject Independent experiment: effects of sequence length (T) on AU intensity recognition rates

.

3.4.3 Discussion: Temporal Analysis

In this section, an HMM-based approach was used to model the time sequence of facial expressions in AU intensity coded DISFA database. We modeled the intensity of 12 AUs and designed both *"subject-dependent"* and *"subject-independent"* models for representing the dependencies among AU intensities. The results show that HMM can model the facial behavioral intensities of a subject very well and in overall get an accuracy about 83% (overall ICC $\approx$ 0.65). In contrast the intra-subject modeling shows that the dynamics of AU intensities and temporal patterns are sharply different among subjects (overall ICC $\approx$ 0.45). This suggests that to have a more accurate automated AU intensity recognizer the relation among different AUs in addition to the temporal pattern dependencies may be helpful to be utilized. In collaboration with other graduate researchers, we investigated the AUs relations and dependencies in automated AU dynamic modeling [96, 169]).

3.5 Posed vs Spontaneous Facial Expression Analysis

In spite of the significant difference between the appearance of spontaneous and posed facial expressions, majority of the previous studies on automatic facial expression analysis, have focused on the posed data. Posed facial expressions are characterized by the exaggerated changes of facial features, however the spontaneous ones have been acknowledged to fundamentally have subtle facial muscle movements with different temporal patterns. For instance Cohn and Schmidt [45], shown that spontaneous smiles, in contrast to the posed smiles, are slow in onset (i.e. rising time), have multiple rises of the mouth corner muscle activation and are accompanied by other facial muscle movements (e.g. cheek raiser).

Reviewing the state-of-the-art literature in the humans affect recognition reveals that there are a few tentative investigations for modeling the posed and spontaneous data. For instance Littlewort et al. [99] compared the facial expressions of pain in posed and spontaneous situations and Valstar et al.[154] discriminated the spontaneous and posed eyebrow actions using temporal dynamics of expressions. However to the best of my knowledge, none of the existing literature specifically had modeled the differences of spontaneous and posed facial expressions (either in term of discrete emotions or discrete AUs). In this regard, for this section of the my dissertation I will focus on comparing the posed and spontaneous data in more detail.

In the last few years, spontaneous facial expressions have attracted several computer vision and artificial intelligent scientists to design and implemented algorithms which compared the reliability of modeling the spontaneous and posed facial actions. Bartlett et al. measured the intensity of action units in genuine and posed facial expressions [24]. They used Support Vector Machine and Ada-boost algorithm for detecting and measuring the dynamics of 20 AUs. They reported the average reliability of 0.63 and 0.30 for predicting the intensity of AUs for the posed and spontaneous facial actions, respectively. The results demonstrate the complexity of spontaneous expressions recognition and low accuracy of automated systems. Valstart and his colleagues [154] used Gentle Boost feature extraction technique integrated with Relevance Vector Machine to model the temporal patterns (i.e. onset, apex and offset) eyebrows muscle movements for posed and spontaneous facial expressions. Their algorithm recognized the class (posed vs spontaneous) brow actions with 90% recognition rates. Recently to facilitate modeling and recognizing genuine expressions, a few spontaneous facial expression databases, such as UNBC Pain Archive [103] and Denver Intensity of Spontaneous Facial Actions (DISFA) [111] have been released.

These research and studies aimed to bring more insight to the area of spontaneous facial expression and determine the main characteristics of facial muscle which allow the computer to model and analyze the spontaneous facial expressions more accurately. The success in this field would improve the efficiency and reliability of human affective state analysis, and creates new opportunity to use the autonomous facial expression recognition systems in real world applications. As it has been repeatedly reported in surveys [], the large number the existing autonomous systems are focused on (i.e. trained and tested) the posed facial expressions; however these exaggerated facial expressions may be significantly different from the genuine expressions there is not an available study, which compared posed and spontaneous facial expressions comprehensively.

Evidence has shown that in overall, posed and spontaneous facial expressions have some intrinsic similarities and differences (e.g. different AU co-occurrence and various dynamics and temporal patterns), but to the best of our knowledge none of the existing studies compared these differences comprehensively. In other words, the un-addressed question is: *In what sense and how the posed facial expressions are differentiable from the spontaneous ones?* Having such knowledge, will definitely help researchers to understand the similarities and differences between posed and spontaneous expressions. This knowledge provide unique insight to designing process for having a reliable autonomous system and transfer the knowledge which has been learned from the posed expressions in the last few decades to the spontaneous category. The lack of such information, attracted me to capture and annotate a diverse set of posed facial actions (for a subset of individuals who had originally participated in DISFA database) and called this new database DISFA+. More details about software I developed to capture the data and also AU intensity annotation and the experimental results I conducted to compare posed and spontaneous data are reported in the following sections.

3.6 DISFA+: A Posed Facial Expressions Database

As presented in Section 3.1, DISFA contains the spontaneous facial expressions of 27 young adults who viewed video clips intended to elicit spontaneous facial expressions. To have a setting which allows us to compare the spontaneous and posed facial actions, we asked a subset of DISFA's participants to rejoin our study again. We recruited nine out of 27 subjects of DISFA and recorded their posed facial actions. These participants have different ethnicities (e.g. Asian, African-American, Caucasian) and their images are shown in Figure 3.9. All of the participants gave us the informed consent for the distribution and use of their video images for research purposes.

3.6.1 Recording Procedure for Posed Expression

To acquire the posed expressions of individuals we designed a software which instructed and guided users to imitate a set of 42 facial actions during the capturing session (some of these facial actions are shown in Figure 3.9). Every subject watched a 3-minute demo to learn how to use the software. For each trial of facial expression imitation, the users were asked to mimic a full dynamic of facial actions (i.e. begin with a neutral face, then proceed to the maximum intensity of expression and finally end with a neutral face).

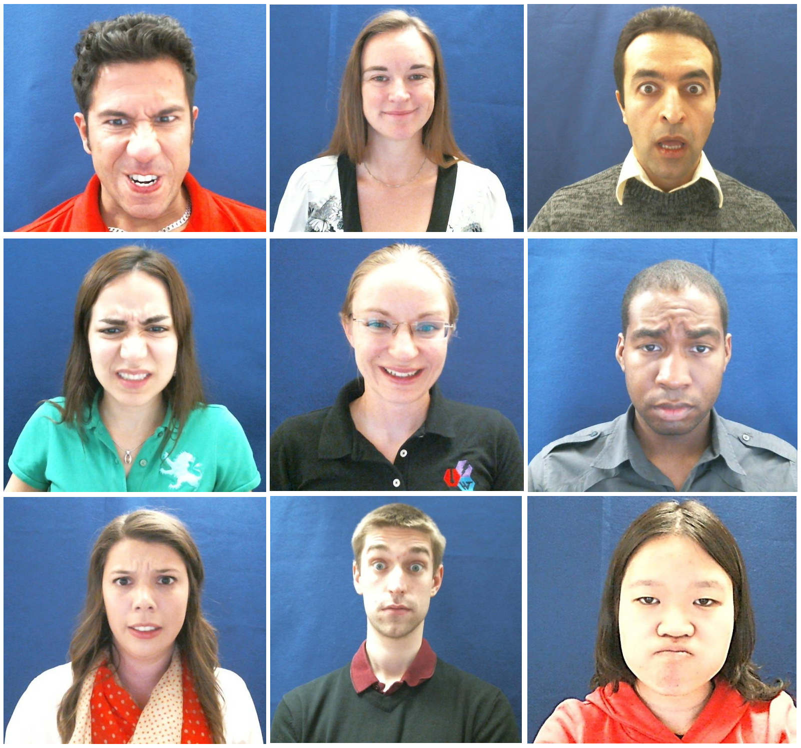


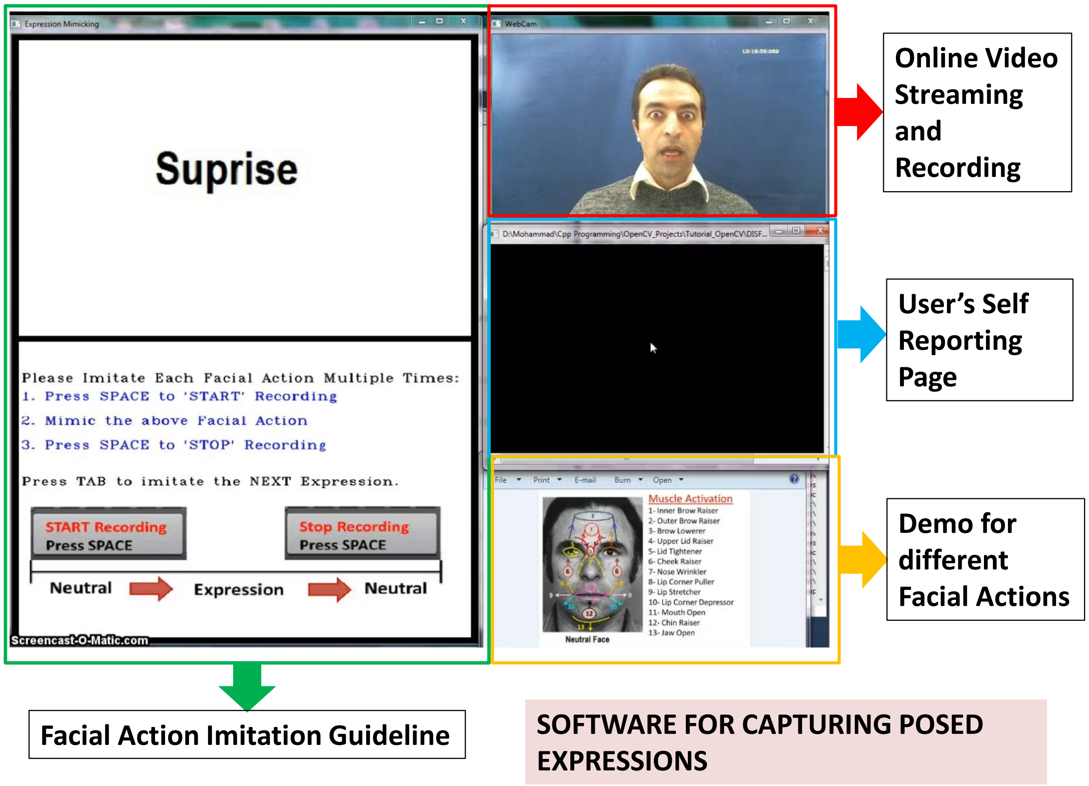
Figure 3.9: Nine subjects of DISFA+ with posed facial expressions.

The users had the option to see their face on an LCD monitor as they mimicked the expression.

Each user was asked to imitate 30 facial actions (i.e. AU combinations) and 12 facial expressions corresponding to the emotional expression (e.g. Surprise, anger, etc.) for multiple times. Meanwhile an HD camera recorded the facial responses of participants with 1280×720 pixel resolution in 20 frames per second. Each individual was instructed to practice a facial action first and after s/he became ready, the subject started recording a few trials for each facial action. Users imitated each facial action few times and after finishing each trial, participants rated (in the range of [0-10]) two factors:

- How difficult it was for him/her to make each facial expression
- How well s/he could make the asked facial expression on the face.

Figure 3.10 illustrates an screen-shot of software while a user showing *'Surprise'* facial action. This software has been developed in C++ and the OpenCV library [32] was used to record and time-stamp every frame of the video.



Please Imitate Each Facial Action Multiple Times:
1. Press SPACE to 'START' Recording
2. Mimic the above Facial Action
3. Press SPACE to 'STOP' Recording

Press TAB to imitate the NEXT Expression.

Muscle Activation
1- Inner Brow Raiser
2- Outer Brow Raiser
3- Brow Lowerer
4- Upper Lid Raiser
5- Lid Tightener
6- Cheek Raiser
7- Nose Wrinkler
8- Lip Corner Puller
9- Lip Stretcher
10- Lip Corner Depressor
11- Mouth Open
12- Chin Raiser
13- Jaw Open

Expression Mimicking
Suprise
START Recording
Press SPACE
Stop Recording
Press SPACE
Neutral
Expression
Neutral
Screencast-O-Matic.com
WebCam
File
Print
E-mail
Burn
Open
Neutral Face
Online Video Streaming and Recording
User's Self Reporting Page
Demo for different Facial Actions
Facial Action Imitation Guideline
SOFTWARE FOR CAPTURING POSED EXPRESSIONS

Figure 3.10: A demo of the designed software for capturing the posed facial expressions of subjects in DISFA+ database.

3.6.2 Manual FACS Coding

DISFA+ provides the dynamics of AU for every frame on a zero (not present) to five (maximum intensity) ordinal scale. In order to annotate the posed AUs I developed a software which allowed the FACS coder to readily annotate and revisit the coded intensity of each AU and save AU labels in a text file. To provide the ground-truth labels for all frames of the captured date in DISFA+, a certified FACS coder annotated the intensity of similar 12 AUs that were coded in DISFA. To evaluate the performance of manual coding, the second FACS certified coder annotated all facial actions of one subject ($\approx$ 12% of total DISFA+

dataset). In total the 6-level intensity levels of twelve AUs of 5,181 frames were re-coded and the inter-observer reliability for all AUs were calculated. The inter-observer reliability of two FACS coders reported in Table 3.7 indicates that for majority of AUs the agreement value between the two coders are high (> 0.70) and confirms the high reliability of the AU annotation in DISFA+.

Table 3.7: AU description and inter-observer reliability of DISFA+.

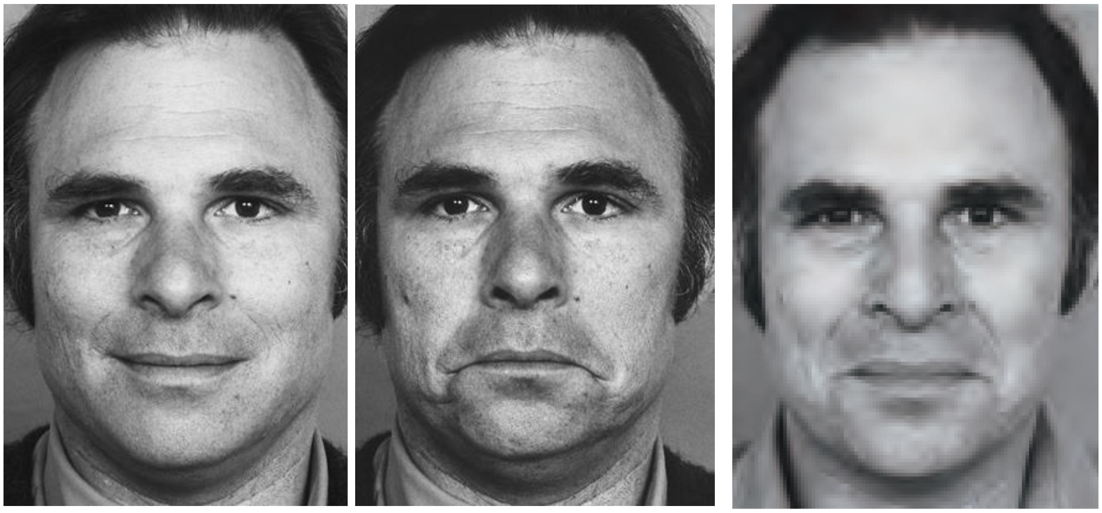
AU	Description	Reliability (ICC)
1	Inner Brow Raiser	0.94
2	Outer Brow Raiser	0.65
4	Brow Lowerer	0.83
5	Upper Lid Raiser	0.75
6	Cheek Raiser	0.80
9	Nose Wrinkler	0.98
12	Lip Corner Puller	0.92
15	Lip Corner Depressor	0.58
17	Chin Raiser	0.88
20	Lip Stretcher	0.61
25	Lips Part	0.98
26	Jaw Drop	0.76

3.7 Comparing Spontaneous and Posed Action Units

The semantic and dynamic relation among facial actions are crucial for understanding and analyzing spontaneous and posed expressions. In fact, the coordination and synchronized spatio-temporal interactions between facial actions produce a meaningful facial expression. Tong et al. [152] and Yongqiang et al [96] employed Dynamic Bayesian Network (DBN) to model the dependencies among AUs for detecting and measuring the intensity of AUs respectively. Their proposed approach demonstrates that using the relation among AUs would be a helpful information to better model facial expressions. In order to compare spontaneous and posed facial expressions, I introduce a few quantitative measures to analyze the AU dynamic relation and temporal patterns in addition to the AU co-occurrence in different facial expression contexts.

3.7.1 Co-presence and co-absence of AUs

Measuring the intensity of AUs in a single frame of video is a challenging task due to the dynamics of facial actions. Moreover, when AUs occur in a combination, they may be non-additive, which means the appearance of an AU in a combination is different from its stand-alone appearance. Fig. 3.11 demonstrates an example of the non-additive effect: when AU12 (lip corner puller) appears alone, the lip corners are pulled up toward the cheekbone (see figure 3.11(a)); however, if AU15 (lip corner depressor) is also becoming active, then the lip corners are somewhat angled down due to the presence of AU15 (see figure 3.11(c)) . The non-additive effect increases the difficulty of recognizing AU individually.



(a) AU12 (b) AU15 (c) AU12+AU15

Figure 3.11: Non-additive effect in an AU combination (a) AU12 occurs alone., (b) AU15 occurs alone., (c) AU12 and AU15 appear together ([65])

As described in the FACS manual [65], the inherent relationships among AUs can provide useful information to better analyze facial expressions. The inherent relations can be summarized as the co-occurrence relations and mutual exclusion relations. The co-occurrence relations characterize the groups of AUs, which oftentimes appear together to show meaningful facial emotions and can also indicate the quality and intensity of an emotion. For instance, in Figure 3.12(a), AU6+AU12+AU25 represents “happiness”, and the two types of disgust faces (Figure 3.12(b) is a sad disgust face: AU9+15+17 and 3.12(c) is a high intensity disgust face with closed eyes: AU9+20+25).

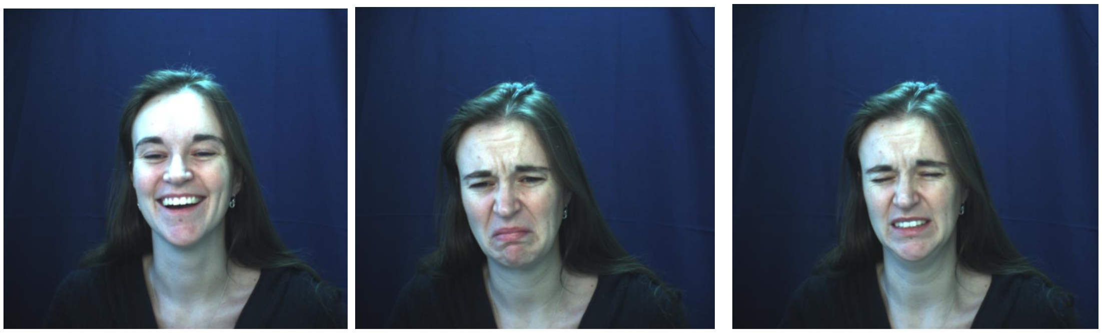

(a) Happy face
(b) Disgust face 1
(c) Disgust face 2

Figure 3.12: Co-occurrence relationships of AUs (a) AU6+AU12+AU25 , (b) AU15 occurs alone., (c) AU12 and AU15 appear together ([65])

On the other hand, considering the alternative rules provided in the FACS manual, some AUs are mutually exclusive since “it may not be possible to anatomically demonstrate AU combinations simultaneously” or “the logic of FACS precludes the scoring of both AUs” [65]. For instance, one cannot perform AU25 (lip part) with AU23 (lip tightener) or AU24 (lip presser) simultaneously. In order to define the co-presence and co-absence of AUs, I used the following definitions [152]:

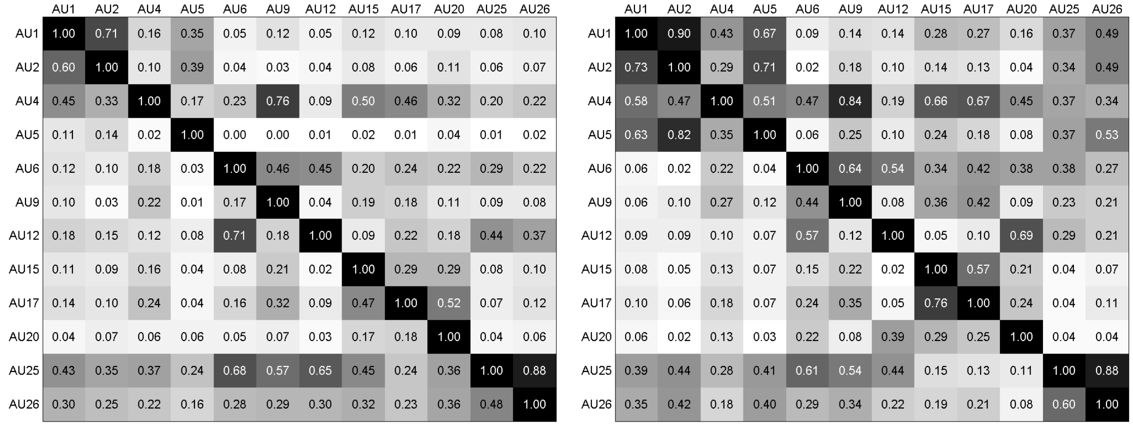
For co-presence (co-occurrence) relations among AUs, a matrix A is defined, where each entry $a_{i,j}$ is the probability of $P(AU_i = 1|AU_j = 1)$ and the pairwise co-occurrence dependency between two AUs is computed as follows:

$$P(AU_i = 1|AU_j = 1) = \frac{N_{AU_i+AU_j}}{N_{AU_j}} \tag{3.7.1}$$

where $N_{AU_i+AU_j}$ is the total number of positive examples of the AU combination $AU_i + AU_j$ regardless of the presence of other AUs in the databases, and N_{AU_j} is the total number of positive examples of AU_j in the databases. Similarly, for the co-absence measurement, a matrix B is defined, where each entry b_{ij} represents the probability $P(AU_i = 0|AU_j = 0)$, and the pairwise co-absence dependency between two AUs is computed as follows:

$$P(AU_i = 0|AU_j = 0) = \frac{M_{AU_i^- + AU_j^-}}{M_{AU_j^-}} \tag{3.7.2}$$

where $M_{AU_i^- + AU_j^-}$ is the total number of frames that neither AU_i nor AU_j occurs, and $M_{AU_j^-}$ is the total number of frames in the databases where AUj is absent. For example as shown in Figure 3.13(a), $P(AU_1 = 1|AU_2 = 1) = 0.71$ means that in the spontaneous facial expression database (DISFA) if $AU2$ is activated, the likelihood of $AU1$ is also contracted is 0.71. In addition, the $P(AU_1 = 1|AU_2 = 1) = 0.90$ in Figure 3.13(b) illustrates that such probability is much higher in posed facial expression data (DISFA+). On the contrary, $P(AU_9 = 1|AU_5 = 1) = 0.01$ means that $AU5$ (eye contraction) and $AU9$ (nose wrinkler) are almost impossible to occur together in the spontaneous facial expressions. In addition, in posed facial expressions some AUs (like AU5) co-occure with few other AUs (e.g. AU1,AU2, AU4, AU25, and AU26) much more frequently than the spontaneous facial expressions.



	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
AU1	1.00	0.71	0.16	0.35	0.05	0.12	0.05	0.12	0.10	0.09	0.08	0.10
AU2	0.60	1.00	0.10	0.39	0.04	0.03	0.04	0.08	0.06	0.11	0.06	0.07
AU4	0.45	0.33	1.00	0.17	0.23	0.76	0.09	0.50	0.46	0.32	0.20	0.22
AU5	0.11	0.14	0.02	1.00	0.00	0.00	0.01	0.02	0.01	0.04	0.01	0.02
AU6	0.12	0.10	0.18	0.03	1.00	0.46	0.45	0.20	0.24	0.22	0.29	0.22
AU9	0.10	0.03	0.22	0.01	0.17	1.00	0.04	0.19	0.18	0.11	0.09	0.08
AU12	0.18	0.15	0.12	0.08	0.71	0.18	1.00	0.09	0.22	0.18	0.44	0.37
AU15	0.11	0.09	0.16	0.04	0.08	0.21	0.02	1.00	0.29	0.29	0.08	0.10
AU17	0.14	0.10	0.24	0.04	0.16	0.32	0.09	0.47	1.00	0.52	0.07	0.12
AU20	0.04	0.07	0.06	0.06	0.05	0.07	0.03	0.17	0.18	1.00	0.04	0.06
AU25	0.43	0.35	0.37	0.24	0.68	0.57	0.65	0.45	0.24	0.36	1.00	0.88
AU26	0.30	0.25	0.22	0.16	0.28	0.29	0.30	0.32	0.23	0.36	0.48	1.00

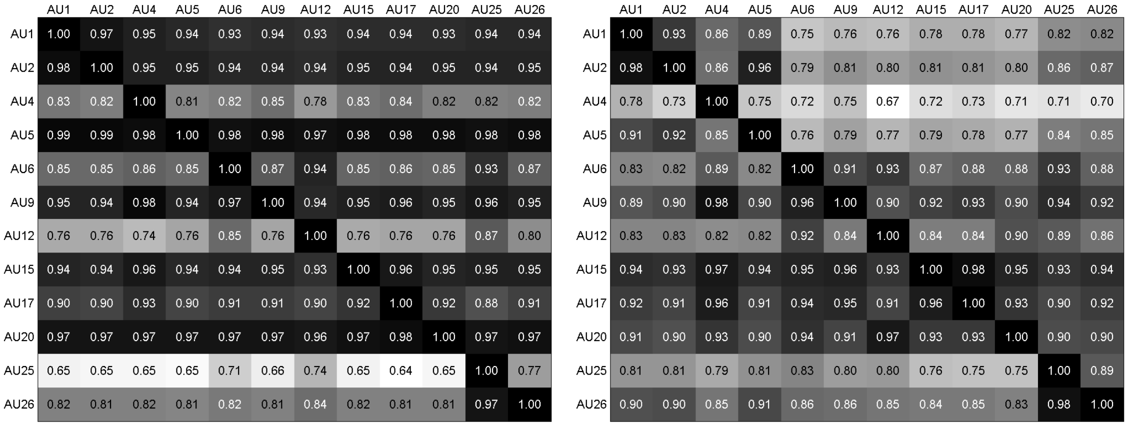
(a) Spontaneous Facial Expressions

	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
AU1	1.00	0.90	0.43	0.67	0.09	0.14	0.14	0.28	0.27	0.16	0.37	0.49
AU2	0.73	1.00	0.29	0.71	0.02	0.18	0.10	0.14	0.13	0.04	0.34	0.49
AU4	0.58	0.47	1.00	0.51	0.47	0.84	0.19	0.66	0.67	0.45	0.37	0.34
AU5	0.63	0.82	0.35	1.00	0.06	0.25	0.10	0.24	0.18	0.08	0.37	0.53
AU6	0.06	0.02	0.22	0.04	1.00	0.64	0.54	0.34	0.42	0.38	0.38	0.27
AU9	0.06	0.10	0.27	0.12	0.44	1.00	0.08	0.36	0.42	0.09	0.23	0.21
AU12	0.09	0.09	0.10	0.07	0.57	0.12	1.00	0.05	0.10	0.69	0.29	0.21
AU15	0.08	0.05	0.13	0.07	0.15	0.22	0.02	1.00	0.57	0.21	0.04	0.07
AU17	0.10	0.06	0.18	0.07	0.24	0.35	0.05	0.76	1.00	0.24	0.04	0.11
AU20	0.06	0.02	0.13	0.03	0.22	0.08	0.39	0.29	0.25	1.00	0.04	0.04
AU25	0.39	0.44	0.28	0.41	0.61	0.54	0.44	0.15	0.13	0.11	1.00	0.88
AU26	0.35	0.42	0.18	0.40	0.29	0.34	0.22	0.19	0.21	0.08	0.60	1.00

(b) Posed Facial Expressions

Figure 3.13: Comparing the co-occurrence of AUs in spontaneous and posed data: (a) spontaneous data (Entire DISFA database) (b) posed data (12 basic expression of DISFA+ database))

To investigate the probability that neither of two AUs occur, I report the co-absence of 12 AUs in both spontaneous and posed facial expressions settings. Figure 3.14 demonstrates that the majority of spontaneous AUs have more tendency to have high co-absence value but very few of them (like AU12 and AU25) may have a low co-absence value.



	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
AU1	1.00	0.97	0.95	0.94	0.93	0.94	0.93	0.94	0.94	0.93	0.94	0.94
AU2	0.98	1.00	0.95	0.95	0.94	0.94	0.94	0.95	0.94	0.95	0.94	0.95
AU4	0.83	0.82	1.00	0.81	0.82	0.85	0.78	0.83	0.84	0.82	0.82	0.82
AU5	0.99	0.99	0.98	1.00	0.98	0.98	0.97	0.98	0.98	0.98	0.98	0.98
AU6	0.85	0.85	0.86	0.85	1.00	0.87	0.94	0.85	0.86	0.85	0.93	0.87
AU9	0.95	0.94	0.98	0.94	0.97	1.00	0.94	0.95	0.96	0.95	0.96	0.95
AU12	0.76	0.76	0.74	0.76	0.85	0.76	1.00	0.76	0.76	0.76	0.87	0.80
AU15	0.94	0.94	0.96	0.94	0.94	0.95	0.93	1.00	0.96	0.95	0.95	0.95
AU17	0.90	0.90	0.93	0.90	0.91	0.91	0.90	0.92	1.00	0.92	0.88	0.91
AU20	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.97	0.98	1.00	0.97	0.97
AU25	0.65	0.65	0.65	0.65	0.71	0.66	0.74	0.65	0.64	0.65	1.00	0.77
AU26	0.82	0.81	0.82	0.81	0.82	0.81	0.84	0.82	0.81	0.81	0.97	1.00

(a) Spontaneous Facial Expressions

	AU1	AU2	AU4	AU5	AU6	AU9	AU12	AU15	AU17	AU20	AU25	AU26
AU1	1.00	0.93	0.86	0.89	0.75	0.76	0.76	0.78	0.78	0.77	0.82	0.82
AU2	0.98	1.00	0.86	0.96	0.79	0.81	0.80	0.81	0.81	0.80	0.86	0.87
AU4	0.78	0.73	1.00	0.75	0.72	0.75	0.67	0.72	0.73	0.71	0.71	0.70
AU5	0.91	0.92	0.85	1.00	0.76	0.79	0.77	0.79	0.78	0.77	0.84	0.85
AU6	0.83	0.82	0.89	0.82	1.00	0.91	0.93	0.87	0.88	0.88	0.93	0.88
AU9	0.89	0.90	0.98	0.90	0.96	1.00	0.90	0.92	0.93	0.90	0.94	0.92
AU12	0.83	0.83	0.82	0.82	0.92	0.84	1.00	0.84	0.84	0.90	0.89	0.86
AU15	0.94	0.93	0.97	0.94	0.95	0.96	0.93	1.00	0.98	0.95	0.93	0.94
AU17	0.92	0.91	0.96	0.91	0.94	0.95	0.91	0.96	1.00	0.93	0.90	0.92
AU20	0.91	0.90	0.93	0.90	0.94	0.91	0.97	0.93	0.93	1.00	0.90	0.90
AU25	0.81	0.81	0.79	0.81	0.83	0.80	0.80	0.76	0.75	0.75	1.00	0.89
AU26	0.90	0.90	0.85	0.91	0.86	0.86	0.85	0.84	0.85	0.83	0.98	1.00

(b) Posed Facial Expressions

Figure 3.14: Comparing the co-absence of AUs in spontaneous and posed data: (a) spontaneous data (Entire DISFA database) (b) posed data (12 basic expression of DISFA+ database))

3.7.2 Temporal characteristics of AUs

One of the important intrinsic characteristic of facial expressions is the temporal patterns of AUs. Therefore, to better describe the dynamics of facial actions, it is helpful to quantitatively measure the dynamics and important time-related features of facial expressions (e.g. rising time, decaying time). Considering that, bellow a few terms are defined to better describe and study the dynamics of facial expressions:

- **Event:** The duration of a facial action that starts and ends with zero intensity and has intensity variations in between.

- **Peak:** the local maximum in the curve of the intensity of a facial action.

- **Valley:** the local minimum in the curve of the intensity of a facial action; (neutral faces and AUs with intensity zero are considered to be a valley)

- **Rising Duration:** the required period of time for an expression to reach from a valley to the next peak.

- **Decaying Duration:** The required period of time for a facial action to decline from a peak to the next valley.

- **Peak Length:** the duration that facial action stays in the peak state.

- **Number of peaks:** the total number of peaks in an event of facial expression.

Figure 3.15 illustrates the dynamics of one facial action event for AU12. This event has three peaks and four valleys and the dashed lines indicate the two middle valleys of the event. One rising time and one decaying duration for AU12 are labeled in the image as well.

Computing temporal patterns of spontaneous and posed facial actions

Using the described temporal terms allowed us to analyze and compare posed and spontaneous facial actions. To report the mean and standard deviation (STD) of the rising and decaying durations, the number of frames in each state were counted. Table 3.8 indicates how rapid each spontaneous facial action may rise or decay in DISFA database. Comparing the mean rising and decaying durations, it can be observed and hypothesized that for every event of spontaneous facial actions, the rising is much faster than the falling time. In order to measure the significance of such hypothesis, the t-test was used. By considering the significance level of 0.05, the t-test confirms that for majority of AUs (except AU2, AU15, and AU17), the rising time is much faster than the decaying time. The **bold** p-values reported in Table 3.8 indicate those AUs, where the null hypothesis of the t-test been accepted.

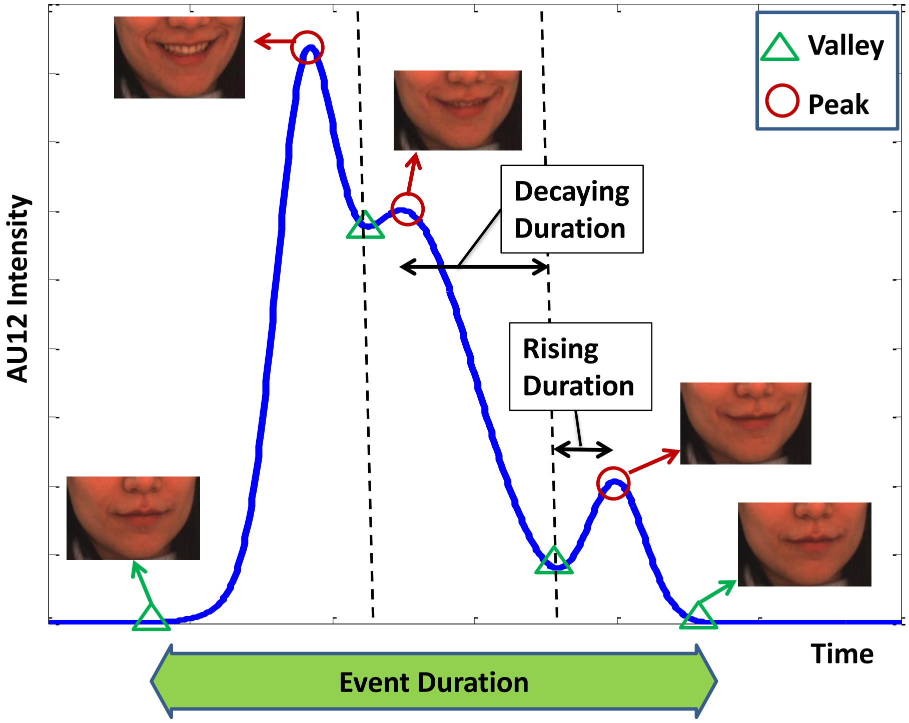


Figure 3.15: Describing temporal features of facial expression in one event of facial expression

Using the same approach for posed data, the rising and decaying duration in DISFA+ database were also compared. Similar to what described for spontaneous data, in posed data, the rising duration is higher than the decaying duration; except for AU6 and AU12 which have very similar rising and decaying times. By considering the significance level of 0.05, and conducting the t-test we can conclude that only six of AUs (i.e. AU1, AU2, AU4, AU5, AU25, AU26) have a significantly higher decaying time than the rising time.

Comparing temporal patterns of spontaneous vs posed facial actions

In order to compare the temporal and dynamic patterns between the spontaneous and posed facial actions, Table 3.10 reports the average of the described features above (e.g. rising time, decaying time, peak length, number of peaks in an event, and length of events) for all the facial action events in both DISFA and DISFA+ databases. The aim of this part of the dissertation is to quantitatively compare different characteristics of spontaneous and posed facial expressions. To evaluate the hypothesis that the spontaneous and posed data have different patterns, the unpaired t-test was used, again. The 0.05 significance level was used to accept/reject this hypothesis.

Table 3.8: Comparing the average rising and decaying time for spontaneous AUs (DISFA Database)

	AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU15**	**AU17**	**AU20**	**AU25**	**AU26**
Rising (Mean)	9.33	12.74	12.50	5.90	17.43	14.04	16.64	14.35	10.59	5.80	22.43	11.70
Rising (STD)	30.06	54.72	35.81	6.37	37.22	27.12	30.19	60.04	40.83	6.06	51.25	25.12
Decaying (Mean)	19.92	20.05	46.19	8.39	24.34	29.83	28.97	21.26	13.67	12.90	39.03	23.19
Decaying (STD)	42.11	44.46	100.98	11.06	41.99	50.17	47.75	43.32	40.87	24.66	109.11	55.86
T-Test(P-Val)	**0.004**	0.131	**0.000**	**0.022**	**0.048**	**0.006**	**0.000**	0.168	0.182	**0.003**	**0.003**	**0.000**

Table 3.9: Comparing the average rising and decaying time for posed AUs (DISFA+ Database)

	AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU15**	**AU17**	**AU20**	**AU25**	**AU26**
Rising (Mean)	7.04	6.09	7.71	6.25	7.24	6.17	7.54	9.52	7.15	6.78	8.06	7.55
Rising (STD)	6.02	4.22	6.66	4.80	5.59	4.83	5.33	5.50	6.15	4.83	8.80	8.25
Decaying (Mean)	10.39	11.20	11.60	12.51	7.17	8.34	7.69	13.44	9.59	6.66	13.70	10.47
Decaying (STD)	11.68	13.55	15.26	14.58	6.30	10.23	6.08	17.91	12.01	6.23	13.69	10.70
T-Test(P-Val)	**0.003**	**0.000**	**0.003**	**0.000**	0.528	**0.083**	0.435	0.150	0.125	0.543	**0.000**	**0.022**

Table 3.10: Comparing spontaneous and posed temporal characteristics (using DISFA and DISFA+ databases respectively)

	AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU15**	**AU17**	**AU20**	**AU25**	**AU26**
Rising Duration (Spont.)	9.33	12.74	12.50	5.90	17.43	14.04	16.64	14.35	10.59	5.80	22.43	11.70
Rising Duration (Posed)	7.04	6.09	7.71	6.25	7.24	6.17	7.54	9.52	7.15	6.78	8.06	7.55
Unpaired T-Test (P-Val)	0.41	0.22	0.11	0.64	**0.02**	**0.04**	**0.01**	0.69	0.59	0.32	**0.00**	0.13
Decaying Duration (Spont.)	19.92	20.05	46.19	8.39	24.34	29.83	28.97	21.26	13.67	12.90	39.03	23.19
Decaying Duration (Posed)	10.39	11.20	11.60	12.51	7.17	8.34	7.69	13.44	9.59	6.66	13.70	10.47
Unpaired T-Test (P-Val)	**0.02**	0.06	**0.00**	**0.02**	**0.00**	**0.00**	**0.00**	0.38	0.53	0.08	**0.01**	**0.03**
Peak Duration (Spont.)	25.42	34.10	29.36	14.67	68.04	45.95	53.50	41.78	24.10	30.61	55.70	37.23
Peak Time (Posed)	25.43	24.52	28.74	23.32	32.91	28.47	28.05	32.56	29.22	26.52	23.44	23.28
Unpaired T-Test (P-Val)	1.00	0.18	0.91	**0.00**	**0.00**	0.06	**0.00**	0.30	0.37	0.46	**0.00**	0.06
# **Peaks in Event (Spont.)**	1.07	1.07	1.22	1.13	1.10	1.12	1.29	1.24	1.07	1.04	1.37	1.14
# **Peaks in Events (Posed)**	1.08	1.08	1.04	1.09	1.07	1.08	1.06	1.06	1.08	1.05	1.04	1.05
Unpaired T-Test (P-Val)	0.88	0.87	**0.01**	0.29	0.50	0.29	**0.00**	**0.01**	0.85	0.78	**0.00**	0.07
Event Duration (Spont.)	56.52	64.19	104.9	55.59	103.98	94.70	124.70	110.90	49.71	46.82	157.1	79.81
Event Duration (Posed)	44.09	47.72	47.69	46.64	51.10	51.27	49.84	52.62	51.88	49.25	47.85	48.65
Unpaired T-Test (P-Val)	0.20	0.13	**0.01**	0.30	**0.00**	**0.00**	**0.00**	**0.00**	0.75	0.66	**0.00**	**0.04**

The results presented in Table 3.10 confirms that:

- On average, the majority of spontaneous AUs (except AU5 and AU20) have slower rising time than the posed AUs, however considering the (two-sided unpaired) t-test results indicate that only four AUs (AU6, AU9, AU12 and AU25) in spontaneous context have significantly separable rising time than the posed one.
- On average, all spontaneous AUs (except AU5) have slower decaying (falling slope) than the posed AUs. Moreover, considering (two-sided unpaired) t-test results confirm that the decaying characteristic for majority of AUs (i.e. AU1, AU4, AU5, AU6, AU9, AU12, AU25, AU26) in spontaneous context is significantly different from the posed one.
- Interestingly, the duration of AUs at peak intensity of facial action is largely variant among different AUs. In other words, although the peak duration of four AUs (AU5, AU6, AU12 and AU25) is significantly distinct in spontaneous and posed facial expressions, overall six of the AUs (i.e. AU1, AU2, AU4, AU15, AU17 and AU20) the average duration that facial muscles been contracted to the maximum intensity are same amount. It also worth mentioning that, some of the spontaneous AUs can be very subtle. For instance, as reported in the Table 3.10, on average would stay in peak intensity for about half a second (i.e. on average 14.67 frames in 20fps videos) and then it decays.
- Some of spontaneous facial actions (like AU12 and AU25) have tendency to have multiple peaks. Figure 3.16 illustrates the histogram of number of peaks for all 12 spontaneous AUs of DISFA database.
- On average the duration of events for posed facial expression is about 50 frames (about 2.5 seconds). However for the spontaneous facial expressions we can observe

that some AUs, such as AU4, AU12, AU15, AU25 have very long duration (over 100 frames i.e. 5 seconds) because they tend to stay on apex for longer time or having multiple peaks.

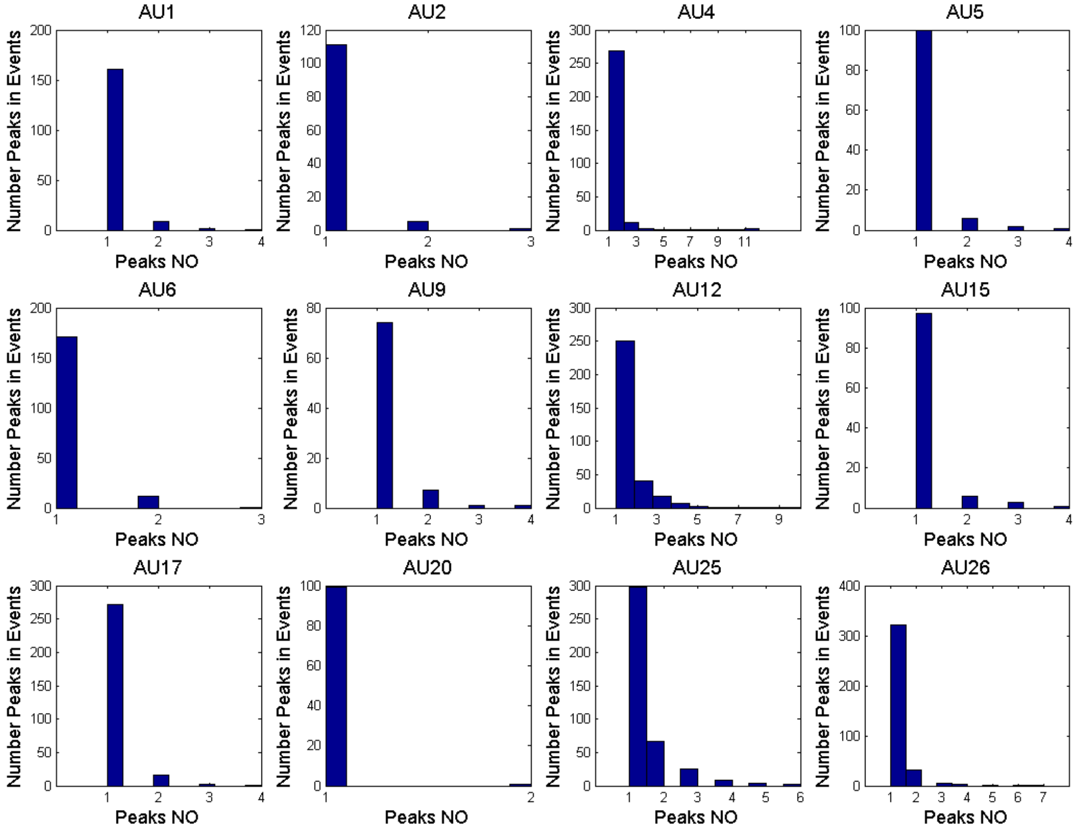


Figure 3.16: Histogram of number of peaks in 12 spontaneous AUs (DISFA)

3.8 Automated Computing of Posed and Spontaneous Facial Actions

Besides reporting the temporal characteristics of AUs in Section 3.7.2, accessing both posed and spontaneous facial expression data allows us to investigate spontaneous and posed data more comprehensively. Bellow are a few questions that interest me to investigate:

1. How reliable and accurate a single FER system can recognize and model the dynamics of posed and spontaneous AUs?
2. Can we extend the knowledge learned from posed data to the spontaneous domain. In other words, how accurate a trained classifier in posed domain can categorize the spontaneous facial actions (and vice versa)?

In order to answer these questions, I designed and developed an automated system for analyzing both posed and spontaneous facial expressions. Therefore both DISFA and DISFA+ were used to train and test the system. Figure 3.17 illustrates few steps to automatically register, represent, and classify facial actions. Section 3.8.1 explains about different bl

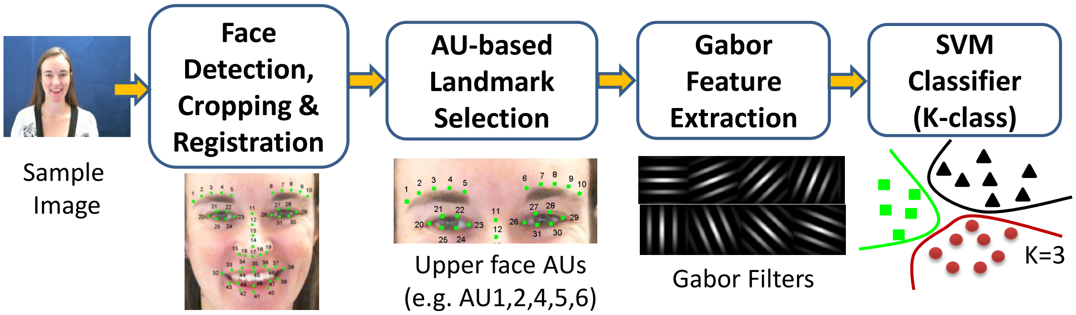

Face Detection, Cropping & Registration
AU-based Landmark Selection
Gabor Feature Extraction
SVM Classifier (K-class)
Sample Image
Upper face AUs (e.g. AU1,2,4,5,6)
Gabor Filters
K=3

Figure 3.17: Proposed block diagram for automated facial expression computing

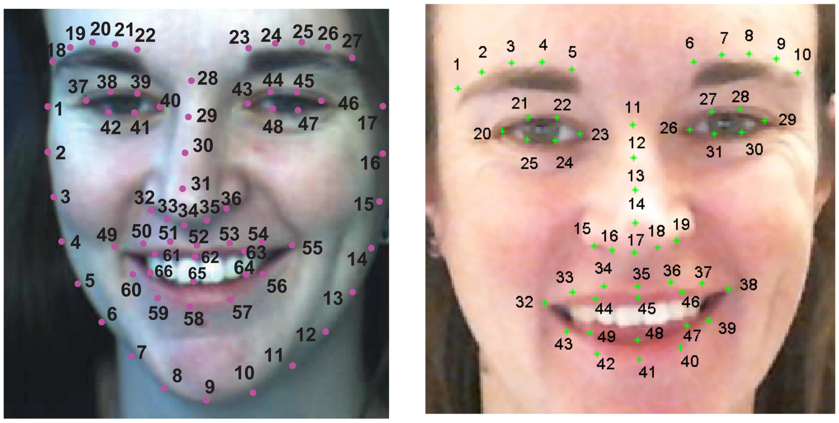
3.8.1 Automated Facial Action Recognition System

This section presents more details about the proposed automated system (shown in Figure 3.17). First, for any given image, I employed IntraFace (a publicly available software [170]) to detect the face regions and extract 49 points around the major facial components (e.g. eyes, eyebrow, lips, etc.). These points were used to register all the images of the database and also extract the localized features. To extract the localized features, 40 Gabor filters (i.e. 5 scales and 8 orientations) were applied at every selected landmark point (see

Table 3.11). For each AU, I used a subset of landmark points, which would best represent the location of the appearance variation for the selected AU. These features then were applied to SVM classifiers. In order to test the results, the Leave-one-subject-out (LOSO) cross validation technique was used.

Study A: Dynamic Computing: posed vs spontaneous

In this section, I use the introduced system in Figure 3.17 to automatically recognize the dynamics of posed AUs. In other words, the goal is to measure the intensity of AUs in a given images. As mentioned in Section 3.1.3, DISFA has 66 AAM landmark point and DISFA+ has 49 landmark points specified by IntraFace. One image sample for both spontaneous and posed data has been shown in Figure 3.18.



(a) DISFA: Registered Joyful Face (b) DISFA+: Registered Posed Happy Face

Figure 3.18: (a) Sixty-six Landmark points extracted by AAM for spontaneous data (b) Forty-nine landmark points specified by IntraFace for posed data.

A set of 300 samples were selected to train an SVM classifier for posed data (RBF kernel with LIBSVM parameter $[-s\ 0\ \ -t\ 2\ \ -c\ 128\ \ -g\ 0.1]$ [36]). In other words, to measure how the learned SVM classifier can recognize the intensity of AUs, LOSO cross validation was used. For both DISFA an DISFA+ datasets, the corresponding 8 subjects (that their AU intensity annotations were ready at the time of writing this dissertation) were used for validation. To compute the intensity of spontaneous facial actions, the described setting (shown in Figure 3.17) for each subject of DISFA was employed to analyze how effectively the proposed automated FER system can compute the dynamics of spontaneous facial actions.

The reported results in Table 3.11 confirm that the trained system has the ability to compute the intensity of posed AUs with higher accuracy and ICC reliability value. However for the spontaneous DISFA database, the trained model can recognize only AU12 and AU25 with ICC values higher than 0.5. This confirms that such method that is capable of measuring the dynamics of posed AUs is not appropriate for modeling the spontaneous data.

Table 3.11: The list of AUs and corresponding landmark point and accuracy for AU intensity measurement.

AU Info		Spontaneous (DISFA Dynamics) results			Posed (DISFA+ Dynamics) results		
AU	**AU Description**	**Lnd. Points**	**Accuracy**	**ICC**	**Lnd. Points**	**Accuracy**	**ICC**
1	Inner Brow Raiser	[21:24 38 39 44 45]	96.58	0.01	[4:7 21 22 27 28]	45.58	0.54
2	Outer Brow Raiser	[18:21,24:27, 37:40 43:46]	99.32	0.00	[1:4 7:10 20:23 26:29]	67.36	0.38
4	Brow Lowerer	[22,23,28]	54.22	0.39	[4:7 11 23 26]	27.83	0.70
5	Upper Lid Raiser	[37:40,43:46]	98.44	0.0	[20:23 26:29]	55.45	0.75
6	Chick Raiser	[37 40:42,43 46:48]	82.22	0.29	[20 23:25 26 29:31]	84.02	0.56
9	Nose Wrinkler	[22 23 28:36 40 43]	98.35	0.0	[5 6 11:19 23 26]	71.90	0.45
12	Lip Corner Puller	[49 50 54 55 60]	52.33	0.56	[32 33 37 38 39 43]	61.36	0.75
15	Lip Corner Depressor	[49 50 54 55 60]	94.89	0.05	[32 33 37 38 39 43]	80.27	0.39
17	Chin Raiser	[57:59 64:66]	58.77	0.11	[40:42 47:49]	53.77	0.20
20	Lip Stretcher	[49 50 54 55 60]	79.62	0.23	[32 33 37 38 39 43]	66.73	0.38
25	Lips Part	[61:66]	40.99	0.64	[44:49]	58.03	0.85
26	Jaw Drop	[49:60]	51.92	0.07	[32:43]	60.14	0.71

Table 3.12: Cross Database ($DISFA \leftrightarrow DISFA+$) Validation: Train on posed data and test on spontaneous date and vice versa.

AU Name	Test on DISFA		Test on DISFA+	
	Accuracy	ICC	Accuracy	ICC
1	66.64	0.05	94.40	0.01
2	99.10	0.00	94.97	0.00
4	16.70	0.33	68.12	0.26
5	85.85	0.17	94.62	0.00
6	29.14	0.30	92.71	0.42
9	95.56	0.05	97.01	0.00
12	67.03	0.54	34.53	0.43
15	79.79	0.07	89.90	0.43
17	58.60	0.05	15.02	0.07
20	79.44	0.15	80.71	0.26
25	33.53	0.35	67.12	0.59
26	34.16	0.19	86.32	0.45

Study B: Cross Database Validation

This section investigates the possibility of training an automated facial expression analyzer on one dataset and apply that to the other one. In other words the objective is to find out if we train a classifier on a posed facial expression, it can be applied to compute the dynamics of spontaneous data (and vice versa). To run the experiments, I extracted the localized Gabor features as explained in Section A.1 and employed SVM with the same set of parameters. Table 3.12 reports the results of the cross database validation. Overall, the results demonstrate that the appearance of facial expressions in posed and spontaneous data are significantly different which cause the automated algorithm having difficulties in recognizing the dynamics of AUs.

3.9 Conclusion

Data drives research. Without adequate data for training and testing classifiers, progress is difficult at best. Most previous databases for automated facial expression analysis were limited to posed facial expressions, holistic labels (e.g., emotion labels), or continuous ratings along a molar dimension (e.g., valence). We present a new, publicly available database of spontaneous facial behavior in response to emotion eliciting videos. Facial behavior is exhaustively annotated using anatomically based descriptions, AUs. The intensity of each AU is annotated on a six-point intensity scale. DISFA thus provides continuous annotation of graded changes in spontaneous facial expression of emotion. To promote research in automated facial expression analysis, DISFA includes 66-facial landmarks for each video image. For each representation, we report benchmarks performance for SVM classifiers for AU intensity on an ordinal scale. In addition, we reported results of an HMM classifier for modeling and computing dynamics and temporal characteristics of facial actions.

In addition to comprehensively investigate the temporal and dynamic characteristics of facial expressions (in both posed and spontaneous domains), we collected and annotated the posed facial actions of few participants in DISFA. This extension is called DISFA+. A FACS coder annotated the intensity of images in DISFA+ and these scores were used to model the temporal patterns of AUs. To compare posed and spontaneous facial expressions, I defined a few terms to describe the temporal patterns of posed and spontaneous data. Comparing rising and decaying times of AUs revealed that majority of spontaneous AUs has slower rising and decaying time than the posed ones. Also an automated FER system was proposed and used to investigate and model the characteristics of facial actions. We believe that both facial databases and presented protocols and benchmark results will be a great asset for the science community and will help researchers to develop and evaluate novel methods for spontaneous facial expression recognition that eventually can be used for real-world HMI systems.

Chapter 4

Autism and Human-Robot Interactions

The second objective of my dissertation is to design, develop, and study robot-based protocols to investigate facial behavior responses of children with autism and applying HRI systems in real-world therapeutic applications. Due to the fact that children with autism show interest toward technology and as the gaze direction and head orientation patterns are important facial behaviors in face-to-face communication I utilized statistical and automated techniques to study, how facial response of children with autism are different compared to typically developing children in HRI systems. In addition, my colleagues and I designed and executed a set of robot-based games to study and foster the socio-behavioral responses of children diagnosed with high-functioning autism.

4.1 Autism Spectrum Disorder

Autism or Autism Spectrum Disorder–ASD is a general term used to describe a spectrum of complex developmental brain disorders causing qualitative impairments in social interaction. Individuals with autism oftentimes have deficits in appropriate verbal and non-verbal responses, including motor control, mimicking facial expressions, and eye gaze regulation. Next section reports some of the prevalence of autism and common challenges that children with ASD, their family members and community encounter.

The Autism and Developmental Disability Monitoring (ADDM) [54] network is an active surveillance system, which provides the statistics related to the prevalence of ASD among children aged 8 years (i.e. the child's parents or guardian should live in one of 11 ADDM sites in USA). Every ASD child in this site has been examined by qualified trained clinicians and had a comprehensive evaluation. If the child meets the definition of one of the ASD conditions (e.g. autistic disorder, pervasive developmental disorder-not otherwise specified (including atypical autism), or Asperger syndrome) then s/he has been included in the study. ADDM provided an comprehensive study [54] and reported some statistics related to children, including the prevalence estimates and characteristics of population of children with ASD. In 2010, [11] provides a list of important prevalence rates and statistics related to ASD and some of them are reported here:

- About 1 in 68 children has been identified with ASD.
- ASD is reported to occur in different racial, ethnic, and socioeconomic groups.
- ASD is almost 5 times more common among boys (1 in 42) than among girls (1 in 189).
- About 1 in 6 children in the USA had a developmental disability in 2006-2008 (ranging from mild disabilities such as speech and language impairments to serious developmental disabilities, such as intellectual disabilities, cerebral palsy, and autism).
- Research has shown that a diagnosis of autism at age 2 can be reliable, valid, and stable.
- On average, children identified with ASD were not diagnosed until after age 4.
- It is estimated to cost at least $17,000 more per year to care for a child with ASD compared to a child without ASD.

- Children and adolescents with ASD had average medical expenditures that exceeded those without ASD by $4,110 to $6,200 per year.
- In addition to medical costs, intensive behavioral interventions for children with ASD cost $40,000 to $60,000 per child per year.

4.2 Social Interaction and Facial Behaviors

Humans learn social communication and interaction skills at a very early stage of their life [94]. Newborn children have an innate ability to respond to surrounding events in the environment, such as detecting and tracking the face [118] and recognizing vocal features of their mother [73]. Through the first year of childhood, children develop several aspects of socio-communicational skills through physical and social interactions with their parents (i.e. caregivers) and their toys [153]. For Typically Developing (TD) children, one of the earliest social skills that the child acquire is to maintain eye contact and shift the gaze direction toward an object of interest. Figure 4.1 demonstrates how these social competencies are developed in the first year of life. As demonstrated in Figure 4.1, eye contact and joint attention are two main skills that are developed and maintained at the very early age and humans employ these skills to better convey their intentions in social environments [94]. Next section describes a few important aspects of gaze direction regulation and joint attention which allow us to facilitate social-verbal interactions.

4.2.1 Gaze Perception in Social Interactions

Human gaze serves as an important cue for not only monitoring attention levels, and emotional states of others, but also for temporally synchronizing human-human interactions [94]. There have been several research related to investigating and analyzing human gaze and the relations of gaze direction with the person's behavior [84]. For instance it has

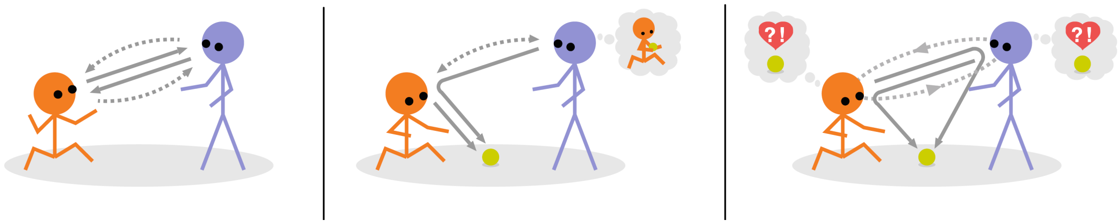


Figure 4.1: Three stages of social development in the first year of life [94]:
-Left *Under 3 months of age:* Child establishes ”eye contact” (the joint action of two individuals looking into each other’s faces) and exchange vocal and facial expressions with the caregiver.
-Middle *3-9 months of age:* Caregiver interprets and responds to the child’s mental states (e.g. desire, pleasure, dislike) and the child gradually learns to predict caregiver’s behaviors.
-Right *Over 9 months of age:* ”Joint attention” (Two people look at the same point or object by means of gaze and/or gesture), empathic understanding and shared action regarding a target, emerges in child-caregiver interaction.

been shown that gaze can significantly affect the quality of conversation and sometimes can reinforce the speech contents. Oftentimes we regulate gaze patterns to initiate, emphasize or terminate our conversation [84]. Therefore, gaze direction and staring pattern can improve the verbal and social aspects of our conversation. For instance, [18] demonstrated the importance of visual cues and gaze patterns for detecting the end of utterance, which makes human feel more comfortable during the interaction.

Studies have shown that appropriate eye contact can improve the quality of conversation, but too much eye gaze can increases the arousal level, which may disturb the dyadic interaction [25]. In 1965, Argyle and Dean [16] developed an *intimacy equilibrium model*, which provides a linkage between gaze an other non-verbal communication cues. This model suggests that one cue can increase the intimacy level (i.e. be closer, look more, smile more often, etc.) of human with others, but will reduce the other cues.

4.2.2 Facial Expression Recognition and Emotional Development

The ability to recognize and mimic facial expressions is important component for conducting social interactions [80]. In a study conducted in 2011 [132], the authors illustrated that facial expressions are developed before birth. They observed that babies in the womb develop a range of facial movements, throughout 24 to 36 weeks of pregnancy [132]. In another interesting study, Galati et al. [75], compared the facial expressions of 10 blind and 10 sighted children (8-11 years old), to find out how humans can perceive and illustrate facial expressions for different emotions. They noticed that both groups (blind and sighted) have similar facial expression responses to the stimuli, although the facial muscle movements of blind children are faster than the sighted peers[75].

Few studies have explored the development of emotion identification during infancy, childhood and adolescent [112, 157]. It has been shown that different factors such as gender, socio-economic status, and verbal ability may also effect on emotional developmental pace [34]. Functional neuro-imaging studies in adults has highlighted the role of the amygdala (aka a group of nuclei that has the primary role for processing, decision making and emotion reactions, which belongs to a part of limbic system) in emotion processing [80].

4.2.3 Socio-communicational Deficits in Individuals with ASD

Using a rich set of emotional cues can help us to better show how we feel. Facial expressions is a powerful nonverbal sign that humans are using involuntary and unconsciously in their social interactions which can convey their moods of an individual. Facial expression, oftentimes co-occur with other non-verbal communications cues, such as gaze responses, body gesture, head orientation, etc. TD individuals are able to use and regulate the intensity level of each of these cues to convey their emotional states to others. In contrast, several individuals with neuro-developmental disorders (e.g. autism) are characterized by dysfunction in emotional and cognitive aspects. Following sections will discuss the major difficulties of children with ASD and existing robot based interaction systems.

Individuals with ASD usually experience verbal and nonverbal communication impairments which weaken their ability to interact with others effectively. Oftentimes, they have deficits in various facets, such as (1) language delay, (2) difficulty in having empathy with their peer and understanding others emotions (i.e. facial expressions recognition.), and more remarkably (3) joint attention (i.e. eye contact and eye gaze attention). Autism is a disorder that appears in infancy [23]. Although there is no single accepted intervention, treatment, or known cure for ASDs, these individual will have more effective treatment if it is diagnosed in early ages. At the first glance at an individual with autism, you may not notice anything odd; however after trying to talk to him/her, you will understand something is definitely not right [77]. S/He may not make eye contact with you and avoid your gaze [77].

Considering the high prevalence rate and high costs of medications and therapeutic sessions, besides the wide range of deficits that each ASD individual may have, the question is *"How can we provide a functional and acceptable treatment to the children with autism?"* As individuals with autism are not fully-engaged and interested in having a social interaction with other humans, in the last decade several investigations have been designed and conducted to recognize and model the behavioral responses of individuals with autism and find out how we can employ technology to diagnose and treat their deficits effectively. Recently, several machine-based (e.g. computers, robots, etc.) technique have been proposed that may alleviate the autism diagnosis process and may provide more efficient and economical therapeutic environment for ASD individuals. Section 4.3 discusses some of the human-robot interaction approaches.

4.3 Literature Review: Autism and HRI

Autism Spectrum Disorder refers to a group of mental developmental disorders which can negatively affect humans social and communicational interactions and ultimately isolates the individuals from the community. As reported by the Center for Disease and Control Prevention 2014, one in 68 American children has been diagnosed with ASD with an approximate annual increase of 10 to 12 percent [55]. According to the DSM-5 diagnostic criteria, individuals with ASD experience deficits in social-emotional reciprocity, verbal and nonverbal communication behaviors, and deficits in development and maintenance of relationships [12]. Characterized as an early brain developmental disorder, ASD can be diagnosed in an individual as young as 2 to 3 years of age [47, 117]. As surveyed in [135], social skills and communication development intervention for children at the earliest age possible is super beneficial and efficient, hence autism community (i.e. parents, caregivers, psychologist, etc.) is actively looking forward to diagnose and treat the deficits of children with ASD at very early stages.

With this increase of prevalence and need for early intervention, various forms of practices have been developed for treating children and young adults with ASD [167]. In the 2014 National Autism Center Report, 27 interventions with well-documented evidence exist for behavioral intervention of ASD [168]. Each of these interventions consists of Applied Behavior Analysis (ABA) techniques using systematic ways of producing replicable procedures [167]. Although the majority of existing interventions are delivered through human-to-human interaction, the most recently accepted intervention utilizes the Technology-aided Instruction, where technology is referred to as any electronic item, equipment, application or virtual network [119]. While technology-based intervention has been a relatively new therapy procedure, it has been growing rapidly due to the recognition that many children with ASD show a collective interest toward technology [133].

With the advancement of robotics technology, investigation on the use of Socially Assistive Robots (SAR) for treatment of ASD has grown recently. Beginning in 1976, use of robots was first reported positively with the use of a remote controlled robot to elicit unprecedented communication with a boy with Autism [163]. Work by [50] in the AuRoRA project further showed how robots could be tools used for Autism therapy. When two children diagnosed with ASD were paired with a robot mediator, the researchers noted increased shared attention between the children rather than when working through a human mediator. In the AuRoRA project, observations were also made of how children with ASD preferred robotized versions of objects such as dolls, trucks, or humans rather than their original forms [50].

Furthermore, studies such as [91] have shown that children with ASD increase responsiveness and interaction with a human when paired with a robot rather than with another human or computer screen. As noted by [61], it is hypothesized that the robots advantage is likely due to its simplified social complexity versus a human but physical presence not possible with computer screen technology. Thus promising results in use of robots for treatment of ASD has promoted further study of clinical applications for interactive robots.

A few quantitative studies have been published that investigate the use of robots for teaching or improvement of social skill. Some of these include Duquette et al. [61] and [162] . Duquette and colleagues [61] found that two low-functioning children diagnosed with ASD interacting with a human mediator were better capable of imitating body movements than those paired with a robot named, Tito. However, Duquette et al. noticed that participants had a heightened interest to their robotic mediator than the human explained by the attracting characteristics of the robot such as lights, color, and movement. Furthermore, use of the robot mediator increased imitation of facial expressions and shared attention between the child pairs when compared to sessions with a human mediator.

Warren and his colleagues [162] investigate the use of feedback and intervention to increase joint attention, the use of gaze or gestures to have the child and robot share attention towards a common target. To trigger such joint attention, the robot in this study, NAO, turned its head towards a target monitor set either to the left or right of the child participant. If the child successfully looked at the target directed by the robot, a technician rewarded the child by playing a clip of a childrens cartoon. Results showed that interest in the robot was sustained and an improvement in performance occurred across four session trials. This improvement was observed through a reduction of prompts needed to elicit the child to direct their gaze towards the proper target.

In 2008, Feil-Seifer and Mataric [72] conducted a robot-based experiment to examine whether or not robot behavior could affect the social behavior of four children with ASD. To test this, they compared two different robot behavior schemes, contingent and random, and their effect on the children with ASD. Through use of the contingent behavior scheme, the robot carried on its rolling and bubble blowing on its own internal schedule, disregarding the behavior of the child. However, using the contingent behavior scheme, large buttons mounted on the robot triggered the bubble blower. The results indicate that when the robot was acting contingently the child was more sociable and could actively engage the children with ASD.

Research by Kim et al. [91] also suggests that the use of robots can serve in improving communication and social skill interventions for children with ASD. This research explored the number of utterances and willingness to converse with a human adult when another confederate was involved. Such confederate could be a robot, adult, or touch screen computer and was randomly paired with each child. Results show that a significant increase in utterances by the child toward the first human adult occurred when the robot was present rather than the other two confederates. A greater amount of utterances was also directed toward the robot rather than toward the other two confederates as well. While studies about the use of robots for social skill improvement are able to make positive observations, research in this area is still exploratory and generally limited to short study durations [91].

The goal of Section 4.6 of dissertation is to design and develop a robot-based protocol to shed light on applying HRI systems in real-world therapeutic and assistive applications. Since children with ASD show interest toward technology, I designed and conducted a set of robot-based games to study and foster the socio-behavioral responses of children diagnosed with high-functioning ASD. The behavioral responses of participants during different phases of this study (i.e. baseline, intervention and follow-up) reveal that overall, a robot-based therapy setting can be a viable approach for helping individuals with autism. In addition, considering the gaze direction modeling allows us to understand how individuals with ASD regulate their facial behaviors differently (compared to typically developing children) when interacting with a robot. Therefore Section 4.4 investigates and models the facial behavioral responses of our participants when interacting with the robot (NAO).

4.4 Protocol A: HRI and Facial Behavior Modeling

Employing socially assistive robots for diagnosis and therapeutic applications is an emerging field in autism research. The main body of the existing literature has focused on investigating how robots can be applied for socially engaging individuals [56, 162]. However, one of the fundamental and un-addressed questions about human-robot interactions is, *How do individuals with autism interact with robots compared to their Typically Developing (TD) peers?*

4.4.1 Methodology

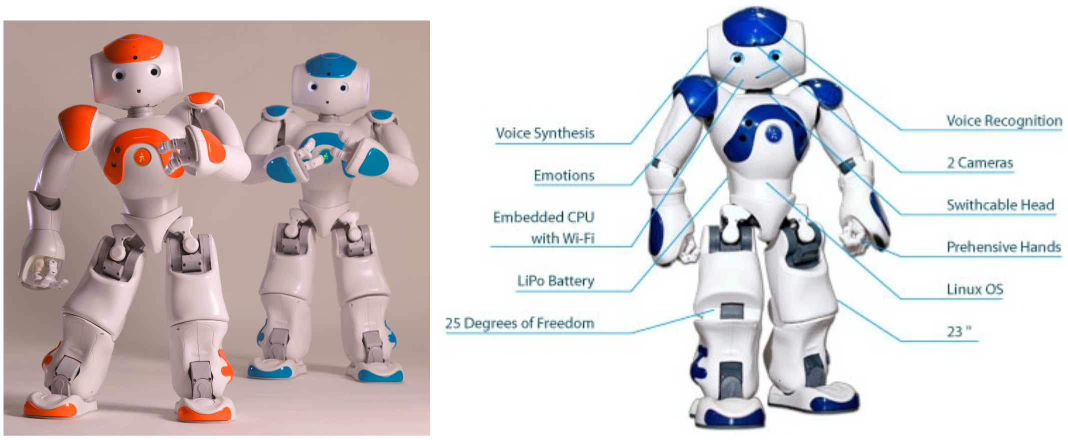
Reviewing the literature in autism research reveals that the majority of existing studies investigated how individuals with autism can regulate their gaze direction either toward different objects or the face regions [16, 144, 162]. However, to the best of our knowledge, there are no studies on how individuals with ASD regulate their gaze direction in conversational contexts. Besides, none of the existing studies investigated the question of: *How do humans (both TD and ASD groups) regulate their gaze responses when interacting with a robot?* Considering the importance of gaze response in social interactions, the main objective of our study is to examine the eye gaze direction of both TD and ASD groups when they interact with a robot. For this segment of dissertation I studies the facial behaviors of ASD and TD children in conversational context with a robot. Following sections introduce the robot, participants and details of human-robot interaction setting.

NAO: A Humanoid Robot

Our test platform consisted of NAO, a commercially available humanoid robot, and a controlling system based on the Choregraphe software for Wizard-of-Oz operation of the robot for interacting with the child in the games. NAO is 58cm (23 inches) tall humanoid robot, with 25 degrees of freedom that was developed by Aldebaran Robotics in France [2]. It has been equipped with an on-board multimedia system including, four microphones for voice recognition, and sound localization, two speakers for text-to-speech synthesis, and two HD cameras (maximum image resolution 1280960) that can be used for on-line streaming or video recording. Figure 4.2 demonstrate a few joints, modules, and postures of NAO. In order to remotely control and monitor NAO, we used *Choregraphe* software with a well-designed user interface which allowed us to control different joints movements, camera capturing and streaming, text-to-speech synthesizing, etc.

Participants

Eleven high-functioning children with ASD (7-17 years old) and four same age TD individuals participated in our study. All of the participants were recruited from Denver,



(a) NAO

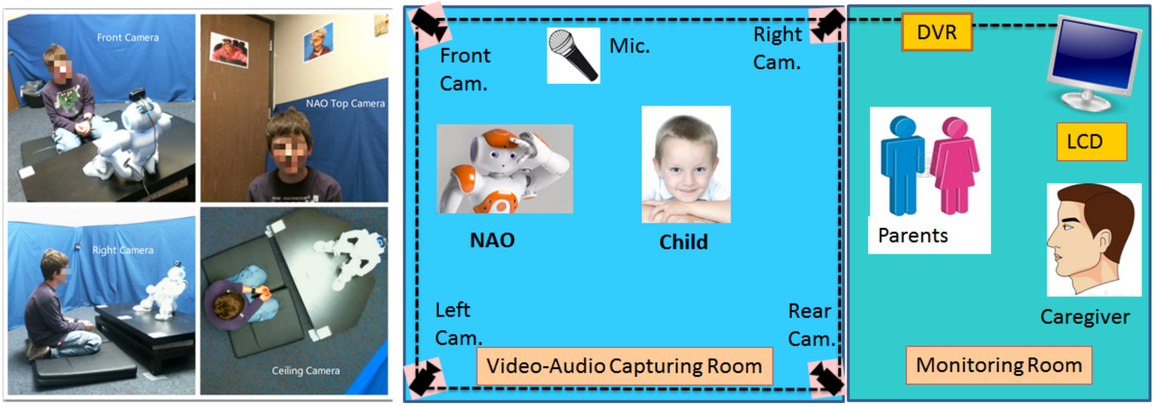
(b) Modules of NAO

Figure 4.2: (a) Two postures-colors of NAO, (b) Embedded modules of NAO and their positioning

CO (USA). The children with autism were recruited from different autism treatment organizations, local autism schools, and families associated with the JFK partners. The parents were asked to provide documentation of their childs ASD diagnosis to participate in the study. The typically developing children were also selected from Colorado schools and we obtained an IRB approval and a signed informed consent form from the parents of both TD and ASD groups.

Room Design and Video Recording Setting

The child-robot interaction was conducted in a $5m \times 5m \times 5m$ room where we captured the video and audio signals of the robot-child interaction. For monitoring purposes, all devices (i.e. cameras, microphones) were connected to a recording system and an LCD screen outside of the room. This allowed the parents, robot operator, and a graduate researcher to observe and listen to the robot-child interaction. During each game we asked the child to be alone with the robot. However, one of the children (i.e. the 6 years old girl) was not comfortable being alone with the NAO (especially for the first session) therefore, her parent stayed with her during the session.



(a) Recording Room (b) Schematic Plan

Figure 4.3: (a) Four camera views of cameras during NAO-child interaction, (b) Schematic plan of capturing and monitoring rooms.

We installed four synchronized cameras on the four corners of the room to capture body gestures and facial behaviors of the child throughout the sessions. The height of the cameras was set to the eye level (about 3 feet high) of a sitting child. NAO sat on a stand that was located at one of the corners of the room to allow the robot to interact with the children at eye level. For annotating and measuring the behavioral and social responses of children, we mostly used the frontal and top NAO cameras shown in Figure 4.4.

4.4.2 Robot-based Games and Conversational Contexts

For this part of our investigation, we analyzed the gaze responses of both TD and ASD groups in listening and speaking contexts while interacting with NAO. We hypothesize various situations of social interactions for both groups of individuals. The first hypothesis tests if the children with ASD have different eye gaze response (i.e. gaze fixation) toward the robot than the TD group. Next, we compare the gaze pattern of TD children toward a robot, in speaking and listening contexts. The third hypothesis evaluates how significantly the gaze of children with ASD is varying in speaking and listening contexts. Our

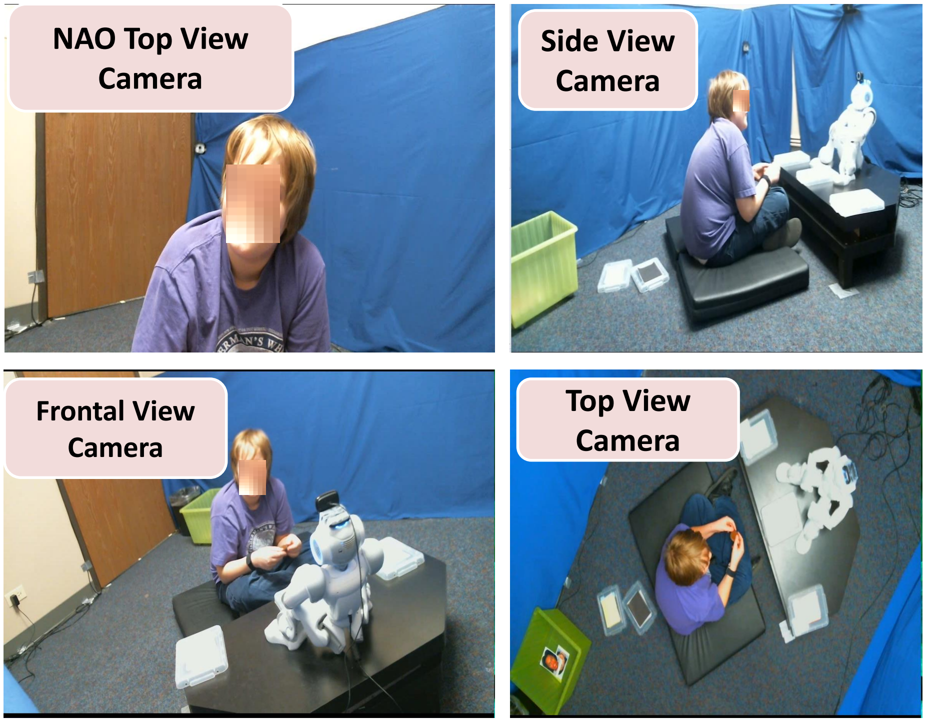


Figure 4.4: Human-robot interaction setting and four-view video recording

investigation of all these hypotheses helps researchers to better model and analyze the gaze responses of both TD and ASD groups for effective human-robot interaction.

We designed a robot-based game to keep the children socially engaged throughout the session. Throughout the game, the participants had the opportunity to demonstrate their social skills, such as verbal capabilities, gaze attention, and social behavioral interaction. In order to have a friendly setting NAO called participants by their name and we used simple language to provide instructions to the child. A few images of different subjects are illustrated in Figure 4.5.

At the beginning of the session, the child was asked to listen to NAO's instructions and respond to the questions, such as "How are you today?" and "What are your favorite movies/books?" The main goal of this segment was to build a friendly relationship between the child and NAO. The robot also asked the child for a hug or a high five in order to

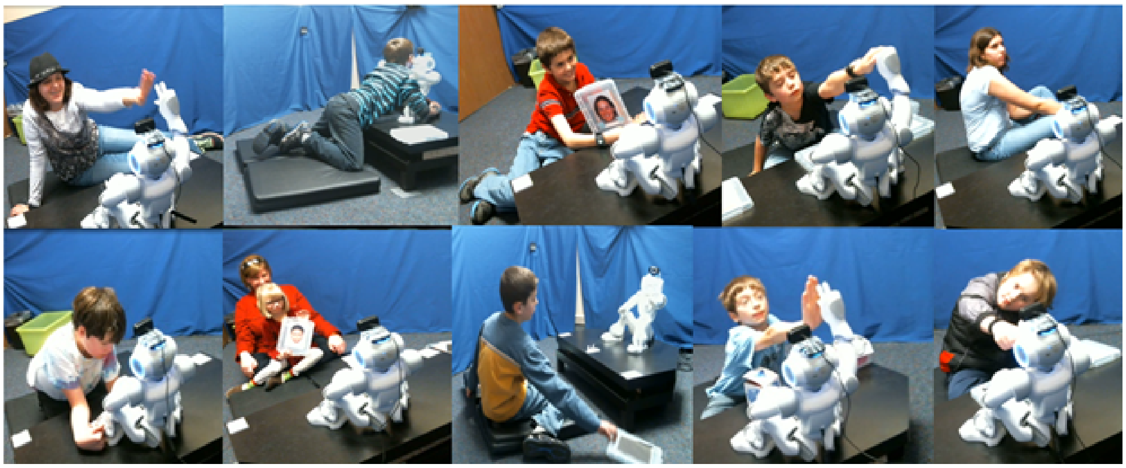


Figure 4.5: Sample images of social interactions of children with NAO

make the participant feel more comfortable and build a positive rapport. Thereafter, NAO initiated a conversation with the participant and asked the child to describe an entertaining activity or pleasant event that s/he had experienced recently. The children oftentimes described movies they had watched or parties or school performances they attended. This part of the game where the participant told the story is called the "*child speaking*" segment. After the child finished telling the story, NAO provided some feedback (e.g. "You did a good job!") and asked the child some questions (e.g. "What do you like about that movie or performance?"). Then NAO began to tell a short story to the child that had exciting content and was easy to understand. This part of the game took approximately two minutes and defined as the "*child listening*" segment. Using the games and the human-robot setting, we were able to engage the child in a communicational and conversational interaction and allowed us to capture and record the child responses throughout the session. I extracted the child speaking and listening video segments and annotated the eye gaze direction (i.e. look at and averted gaze) of the children in those video footages. The gaze annotations have been employed to model and discover the similarities and differences of gaze responses of the TD and ASD groups as they communicate with the robot. The gaze annotation approach, the hypotheses, and achieve results are presented in the following section.

4.4.3 Gaze Directions Coding

The eye gaze responses of all participants in the collected videos (totally 453,000 video frames) were manually annotated using the Continuous Measurement System (CMS) software [114]. A binary coding scheme was used where the averted gazes from NAO were labeled as “0” and the directed gazes at NAO were labeled as “1” (see Figure 4.6). The binary gaze labels were used to measure two important gaze characteristics, 1) Gaze Fixation, and 2) Gaze Shifting Rate as illustrated in Figure 5 and defined below:

- “**Gaze Fixation**” is the percentage of the time that an individual looks at NAO (i.e. counts the number of the frames with label “1” divided by the total number of frames multiplied by 100).

- “**Gaze Shifting Rate**” is the pace of the gaze direction changing from looking at to looking away from NAO and vice versa (i.e. counts the number of changes in the binary gaze labels in the entire length of the video multiplied by 100).

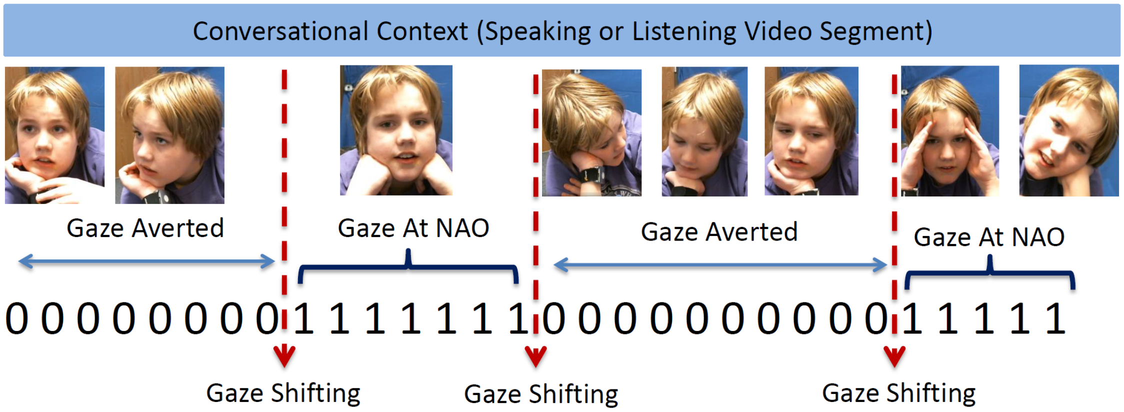


Figure 4.6: Gaze direction annotation: gaze fixation and shifting indices

Throughout the robot interaction games, the child interacted with NAO and we segmented the gaze responses of every participant into two partitions: speaking (“child speaking”) and listening (“child listening”). Then the gaze fixation and shifting rate of each participant were calculated. Using the annotated eye gaze labels, two sets of experiments were conducted. First the between-group (TD vs ASD) gaze responses in a conversational context were compared. In the second experiment, the dependence of gaze patterns on the child's conversational role has been examined. In other words, the study aimed to discover how the gaze responses of each group is linked to the conversational roles.

4.4.4 Experimental Results: Between-Group vs Within-group Gaze Responses

Table 4.1 shows the average and standard deviation of gaze responses in the conversational context for both TD and ASD groups. The results demonstrate that TD individuals (on average) are looking at NAO for more than 71% of the time as they are listening to the robot; however for the speaking context they only spend about 42% of the time looking at NAO. These results are comparable with the previously reported gaze fixation in human-human interaction [16]. In other words, in both human-human and human-robot interactions, the TD individuals are looking nearly twice as much to their interlocutor partner while listening than speaking. Similarly, children with ASD on average, spend more time looking at NAO when they are listening than speaking (61% as opposed to 52% respectively). Although the between-group (TD vs ASD) gaze fixation duration is different, the pattern is consistent and both groups spend more time looking at NAO when they listen. This evidence confirms that on average children with autism have different gaze fixation duration than their TD peers; however for the conversational contexts both groups have the comparable gaze pattern while interacting with the robots.

To ascertain the similarities and differences of gaze fixation patterns of ASD and TD individuals, we statistically compared the eye gaze responses for both contexts. In the Listening Context, the independent (one-tailed) t-tests illustrate that the gaze fixation of children

Table 4.1: Overall Gaze fixation and shifting of TD and ASD group in conversational contexts.

Gaze pattern &	**Typically Developed (TD)**		**Autism Spectrum Disorder(ASD)**	
Contexts	**Fixation(%)**(STD)	**Shifting(%)** (STD)	**Fixation(%)** (STD)	**Shifting(%)** (STD)
Child Listening	71.89 (8.40)	0.02 (0.01)	61.88 (17.74)	0.04 (0.02)
Child Speaking	42.17 (10.07)	0.04 (0.03)	52.19 (30.68)	0.04 (0.03)

with autism is not significantly different from their TD peers ($t(11 + 4 - 2) = -1.47, p = 0.083$). Similarly in the Speaking Context, no significant differences are found on the gaze fixation difference between the ASD and TD children ($t(11 + 4 - 2) = 0.95, p = 0.179$) . These results ratify that, during the robot interactions, both groups of children are showing similar eye gaze responses. This evidence can be critical for between-group clinical analysis, which express that the robot interaction can be considered as a potentially viable treatment agent for children with ASD, because they are looking at the robot with similar gaze patterns as the TD individuals are functioning. Furthermore, another question I addressed was *How do the within-group gaze responses vary as the conversation role is changed?* In other words, *How significantly the gaze response of TD individuals is distinguishable when they are speaking than listening (likewise for the ASD group)?*

Results show that the gaze fixation duration of TD children in the listening context toward NAO is more established than when they are speaking. As expected, TD children tend to look at NAO for a longer period of time while they are listening rather than speaking. This can be as a result of, while TD individuals are speaking they need to break and reinitiate the eye contact to better convey and interact with the interlocutor partner, and they follow very similar patterns as they interact with robots. The paired t-test results show a significant difference of the gaze patterns of TD children as the conversational role is changed ($t(4 - 1) = 4.25, p = 0.011$). However by conducting the same test for the ASD group it has been observed that the gaze fixation of children with autism does not vary significantly as the interlocutor role is changed ($t(11 - 1) = 1.08, p = 0.152$).

4.4.5 Automated Eye Gaze Direction Modeling Using Markov Models

This section will utilized an automated modeling approach to address: “How the gaze responses of individuals with ASD differ from that of Typically Developing (TD) peers when interacting with a robot?” Therefore it reports the differences of gaze responses of TD and ASD group in two conversational contexts: *Speaking* versus *Listening*. Variable-order Markov Model (VMM) was used to discover the temporal gaze directional patterns of ASD and TD groups.

4.4.6 Markov-based Modeling and Gaze Analysis

Markov modeling is a powerful stochastic technique that is widely used to represent and model time-sequential data. The learning ability of the Markovian models (e.g. Hidden Markov Model-HMM) has inspired several researchers to apply them for different data analysis and machine learning applications, such as automatic speech recognition [87] and human activity recognition [171].

Hidden Markov model is the most common Markovian model which contains a set of hidden states (X_i) that produce a symbol (Y_i) at each time. In the HMM algorithm, the output of the model only depends on the current state of the system so the previously observed symbols would not have any effects on the output (i.e. memoryless system). However, there are some other Markov modeling techniques that consider the current and previous states of a system to specify the output (e.g. Variable-order Markov Model).

As the gaze direction of an individual can be highly correlated with the history of gaze direction responses, this section utilizes VMM to evaluate the effect of gaze direction history on modeling the gaze of both TD and ASD groups. Therefore, we trained two VMMs to describe the gaze responses of children. These two models are utilized in analysis of the gaze responses of each group in speaking and listening social contexts.

Variable-order Markov Model

In contrast to the Markov chain models where each variable in a sequence depends on a fixed number of random variables, in VMM the number of conditioning random variables may vary based on the observed sequence. This observed sequence is often called the "context". The main advantage of VMM models is that the order may vary for different sequences, depending on their statistics. As a result, VMM produces a probability "tree" with the null space at the root and the element of the sequence as the leaves of the tree. Figure 4.7 illustrates a second-order VMM tree with five alphabets $\{a, b, c, d, r\}$. Considering the training sequence q_1, this tree can represent the probability of each letter, given the current context of the system.

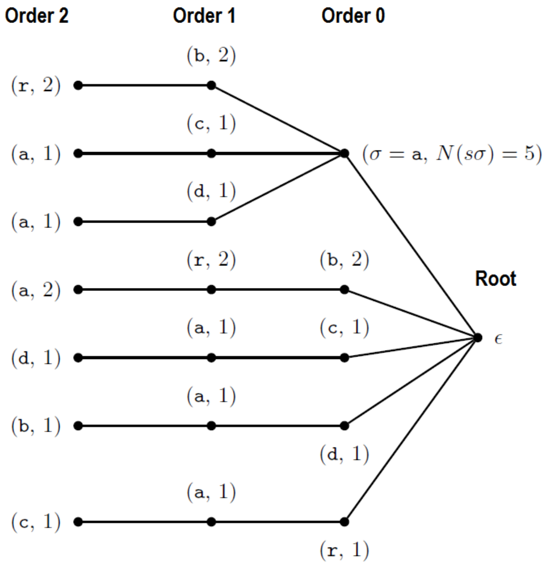


Figure 4.7: The VMM tree constructed using the training sequence $q^1 = abracadabra$ by PPM approach [26].

In order to see how the VMM algorithm is trained, let us assume Σ is a finite alphabet (e.g. $\Sigma = \{A, B, ..., P\}$). Given a training sequence $q^n = q_1 q_2 \cdots q_n$, where $q_i \in \Sigma$, VMM aims to learn a model $\hat{P}$, that can assign the probability of observing any sequence of symbols. Mathematically speaking [26], for any context $s \in \Sigma^*$, and a symbol $\sigma \in \Sigma$, the model generates a conditional probability distribution for $\hat{P}(\sigma|s)$. In VMM the prediction stage utilizes *average log-loss* $L(\hat{P}, x_1^T)$ of $\hat{P}(.|.)$ with respect to the test sequence $x^T = x_1 x_2 \cdots x_T$ (T is the length of observed sequence),

$$L(\hat{P}, x^T) = -\frac{1}{T} \sum_{i=1}^{T} log_2 \, \hat{P}(x_i | x_1, \cdots, x_{i-1}) \tag{4.4.1}$$

There are several different VMM methods that can be applied for classifying a sequence of data [26]. *Prediction by Partial Match* (PPM) is an algorithm that has a robust and reliable functionality. PPM is a finite-context statistical modeling technique that can be interpreted as a mixture of several fixed-order context models of order k [38]. In other words, PPM is a combination of fixed-order context models with different values of k, ranging from zero to a pre-determined maximum (D). For each model, PPM keeps track of the length-k sequence of all characters that have been observed so far in the training sequence.

In addition, PPM can handle the zero frequency problem using *escape* mechanism [26]. In the escape mechanism, the goal is to determine the probability of unseen sub-sequence of symbols after the context s has been seen in the training sequence. In other words for each context s of length $k \leq D$, the probability of $\hat{P}_k(Esc|s)$ is allocated for all symbols that have not appeared after the context s. For instance in Figure 4.7, the probability $\hat{P}(d|ra) = \hat{P}(escape|ra).\hat{P}(d|a) = \frac{1}{2} \cdot \frac{1}{3+4} = 0.07$. More details about training the VMM model is described in [26].

Experimental Results: Automated Gaze Modeling using VMM

Markov models specify a probabilistic output for a given context, therefore training a model with binary letters can be a limiting factor as it can only provide two possible states for each discrete step. In order to code the gaze direction, we used a “phrase alphabet” whose “letters” represent binary sequences of a certain length. For this scheme, we defined the parameter “T” that defines the length of a letter such that each letter corresponds to one of the 2^T possible binary combinations. In a simplistic example where T = 4 (therefore using 16 letters), the sequence 0000 would be re-coded as A, and 1111 would be re-coded as P. Therefore for a given data string, it would be broken down into sequences of four frames which would be translated to their corresponding letters. In this way, the data string for each subject was translated from its original form into a sequence of representative letters. In our experiments we selected 4 consecutive frames to assign $2^T = 16$ symbols (i.e. $\{$’0000’, ... , ’1111’$\}$ or equivalently $\Sigma = [A, B, ..., P]$). Figure 4.8 demonstrates a sequence of gaze labels and the corresponding phrase alphabets for a set of consecutive video frames.

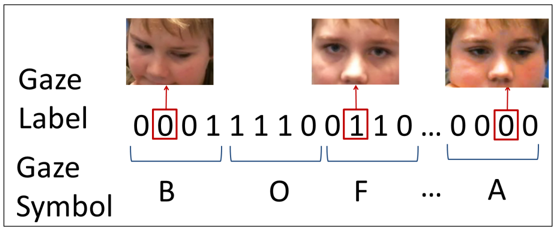


Figure 4.8: Gaze labels (gaze at NAO (1) vs gaze averted (0)) and the corresponding gaze symbols representation ($n = 4$).

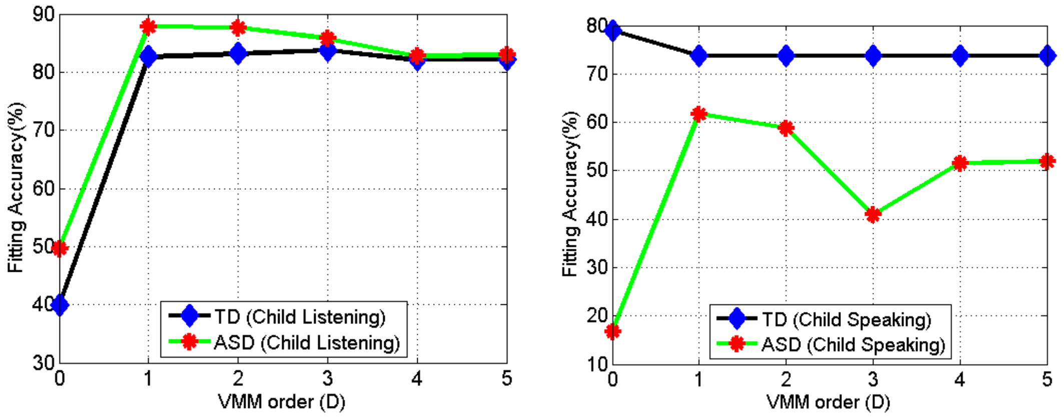
In order to recognize the gaze responses of the TD and ASD groups, we trained one model for the TD group and another model for the ASD group. Therefore, for any given test sample, using Equation 4.4.1 we can measure the similarity of the given test sequence to both models. The test sample is assigned to a class that has the higher similarity value. We measured the overall classification rates and reported the gaze recognition rate of VMM with different orders. Below we explain the proposed algorithm in more detail.

After defining a sequence of gaze letters, we trained the VMM to represent the gaze patterns of the TD and ASD groups. To learn the statistical probability of each alphabet in VMM, we applied the PPM algorithm, described in Section 4.4.6. To find out how the history of gaze direction would play a role in modeling the gaze response of individuals, various VMMs with different orders have been trained and tested. We employed VMM with order zero (memoryless system) as well as the higher order Markov models (system with memory) to study the influence of gaze direction history in categorizing the gaze patterns.

We trained two separate VMM models (one for the ASD and one for the TD group) that describe the gaze patterns of individuals in the ASD group and the TD group. We used Leave One Subject Out (LOSO) cross validation technique to train a Markovian model for each group. In other words, for training the ASD and TD VMM models, we used all the data from all the subjects but one. The excluded subject was used for testing how the trained models for the ASD and TD groups can represent the gaze of each class. Hence, the similarity of the test sample is measured with respect to both the ASD and TD models and the sample is assigned to a model (class) which has the higher similarity value. The same procedure will be applied to all the participants (both ASD and TD) and we reported the average recognition rates for each group.

To evaluate the efficiency of each model, we conducted a log-likelihood similarity measurement (Equation 4.4.1), which specifies the probability of occurrence for the given sequence sample for both ASD and TD models. Using the log-likelihood measure allows us to assign the test sequence to a model that has the higher similarity value. We repeated this procedure for every sequence, and the overall model fitting accuracy for both *Listening* and *Speaking* segments were reported. To measure the importance of history of gaze responses, different orders of VMM (i.e. $D = \{0, ..., 5\}$) were studied. Figure 4.9 illustrates the effects of VMM order on classifying the gaze responses.

The results reported in Table 4.2 demonstrate that, VMM with order one ($D = 1$) can produce the best accuracy for the ASD group. For the TD group, in the child speaking segment the VMM with order one ($D = 1$) and for the speaking context VMM with order



(a) Listening Context

(b) Speaking

Figure 4.9: Comparing Gaze Patterns of TD and ASD groups: (a) Listening Context, (b) Speaking Context.

three ($D = 3$) provides the best results. These results can help us to discriminate between the gaze patterns of ASD and TD groups in different conversational contexts. The reported values confirm that in the "speaking context", the memoryless model ($D = 0$) has higher accuracy for modeling the gaze responses of the TD group, although the VMM with order one ($D = 1$) has better performance for the ASD group. In addition, for the "listening context" the gaze patterns of TD modeled well by higher order VMM ($D = 3$), but for the ASD individuals, VMM with order one ($D = 1$) can represent them accurately.

Table 4.2: ASD and TD Gaze modeling using VMM

Social Context of the Game	**VMM Results** - Fitting Rates					
	D=0	**D=1**	**D=2**	**D=3**	**D=4**	**D=5**
TD (Speaking)	78.95	73.68	73.68	73.68	73.68	73.68
TD (Listening)	39.89	82.58	83.15	83.71	82.02	82.02
ASD (Speaking)	16.81	61.74	58.84	40.87	51.59	51.88
ASD (Listening)	49.64	87.81	87.66	85.78	82.80	82.95

Comparing the gaze responses of ASD and TD individuals in conversational contexts confirms that, the gaze responses vary between the two groups. In other words, different orders of VMM need to be used for modeling the gaze of ASD and TD groups. More specifically, as the TD children are speaking, the memoryless model can represent their gaze patterns with high accuracy (78%). However, when the TD children are listening to NAO, higher orders of VMM ($D = 3$) would best describe the gaze responses. This reveals that as TD individuals are speaking or listening their gaze patterns vary depending on their role in the conversation. For the ASD group, we observed that for both speaking and listening contexts, VMM with order one ($D = 1$) can model the gaze responses well. The results confirm the gaze responses of the ASD group during the speaking and listening situations is not varying significantly and the same order of VMM can model both conversational situations.

4.4.7 Discussion: Eye Gaze Modeling in HRI system

Several studies have shown that individuals with autism are more engaged and demonstrate more positive responses when interacting with a robot rather than a human [51, 140]. However, the intersection of autism treatment research and the engineering (robotic) field has some challenges and limitations, such as different methodologies and objectives. For instance, researchers from the clinical field in autism focus on the treatment and therapeutic aspects. However, due to the existing advancement in both fields it is an interesting era to utilize new technologies for the therapeutic and treatment purposes of teaching social and behavioral concepts to individuals with ASD more effectively.

In the last decade considerable attention has been given to the physical appearance and type (i.e. humanoid vs non-humanoid) of the robot in the effectiveness of engaging individuals with ASD [56]. These studies have helped researchers realize that individuals with autism can understand the physical world more fully and vigorously than the social world [92]. In addition, individuals with autism are more responsive to the administrative feedback via technology rather than a human agent because they are interested in interacting with electronic or robotic components [134]. However these studies lack the emphasis on the best way to integrate the robots into the therapy sessions and how the robots can best play a role in the treatment applications. We investigate the gaze responses of ASD and TD children interacting with a robot in the social contexts and statistically compares the gaze fixation patterns of both groups in conversational context. Our study can help therapists and robotics scientists to design the next generations of games and robots for better understanding and treating individuals with autism.

In our pilot study, we employed NAO which has a great set of human-like features, but a static face. The lack of changeability on the face of NAO has some advantages and disadvantageous, but generally speaking the static and non-emotive face of NAO may results in better engagement of children with the robot. The results of our study confirm that concerning the between-group (ASD vs TD) case, both ASD and TD children demonstrate a similar gaze fixation pattern toward the robot in conversational contexts (i.e. speaking and listening). In other words, both groups spend more time fixing their gaze toward the robot when listening than speaking. Hence we can conclude that NAO is a viable agent for children with ASD for practicing social interactions and engagements.

The similar gaze patterns of ASD and TD individuals toward the robot is clearly a promising factor from the treatment and therapeutic perspective, which demonstrate that individuals with ASD found NAO (a robot with static face and no facial expression) engaging enough for having a conversation with and spend over 50% of their time looking at the robot. However, an unaddressed question still remains: "whether individuals with autism hold the same gaze engagement patterns when the robot has a dynamic emotive face with various facial expressions or not?" More specifically, when a robot uses speech, facial expressions, and other social cues, how do the gaze responses of children with autism alter or divert from the face of the robot. Specifically it would be worthwhile investigating what biometric signals bother individuals with ASD looking at the interlocutor partner if any. Does facial expression, gaze direction, speech intonation or even a combination of these make the social interaction with a human partner more complex and less engaging or are there some other factors in human-human interactions that make it difficult for the ASD group to socialize with human partners.

In addition Section 4.4.6 automatically investigated and modeled the gaze behaviors of ASD and TD children while socially interacting with a humanoid robot (NAO). In this study, Variable-order Markov Models were used to represent and temporally model the dynamics of eye gaze of ASD and TD groups. The results reveal that the gaze responses of the TD individuals in speaking and listening contexts, can be best modeled by VMM with order zero and three, respectively. As we expected, the result show that the temporal gaze patterns of typically developed children are varying when the role in the conversational context is changed. However for the ASD individuals for both conversational contexts the VMM with order one could best fit the data. Overall, the results conclude that VMM is a powerful technique to model different gaze responses of TD and ASD individuals in speaking and listening contexts. Following sections will introduce about VMM and the experimental setup in more details.

Furthermore, as “Autism” is a spectrum of disorders that describes a set of developmental difficulties, another important area to investigate is: *“Which types of individuals with ASD can best take advantage of using technologies and robots?”* In our study we recruited 11 children with high functioning autism to interact with NAO. We realized that even in the same game with a similar set of questions, individuals with high functioning autism varying verbal and gaze response durations (see Figure 7). Although the results of our research study show some of the similarities and differences between the gaze responses of ASD and TD children, further investigation is required with a larger population size, a greater number of sessions and different categories of ASD.

4.5 Head direction and social information

In our daily social and communicational interaction there are several associations between head motions and emotional states. We raise our head in pride and lower it in sorrow [78]. Literature indicates that there is a close relation between head motion, facial expressions of emotion and attention level. For instance, when head pitches down (and often to the side) is a cue for intensifying smile or embarrassment[78]. Head motion can also regulate the humans social interaction. Head turns or nods can be used for turn-taking for other interlocutor partners [60]. Head motions are very reflective of individuals personality. It is also shown that the frequency of head actions are greater in women than men [29].

Although eyes and gaze direction are important cues for measuring the attention level of an individual, but other social responses such as head orientation and body posture, or even pointing gesture can provide useful social information [95]. In an investigation conducted in 1992, Perrett et al [129] hypothesized *"how information from eyes, head and body might be combined in order to detect the attention of human"*. Their study indicates that information from gaze can override that from head cues, and the information from head cues overrides the body cues. It is also worth mentioning that based on the study in early nineteenth century (cited in [33]) the perception of gaze is largely influenced by head direction. For instance in Figure 4.10, the face on the right seems to be gazing directly at the viewer, whereas the face on the left appears to be looking slightly to the viewer's right. However, by covering the lower portions of each face, you can satisfy yourself that the eye regions of both are, in fact, identical. In 1967, Cline claimed that head orientation and eye gaze direction interact in a way that the direction of attention is perceived as falling somewhere between these two positions [39].

The goal of this section of my dissertation is to find out the point of attention for children with ASD when interacting with NAO. Tracking the head posture provides useful

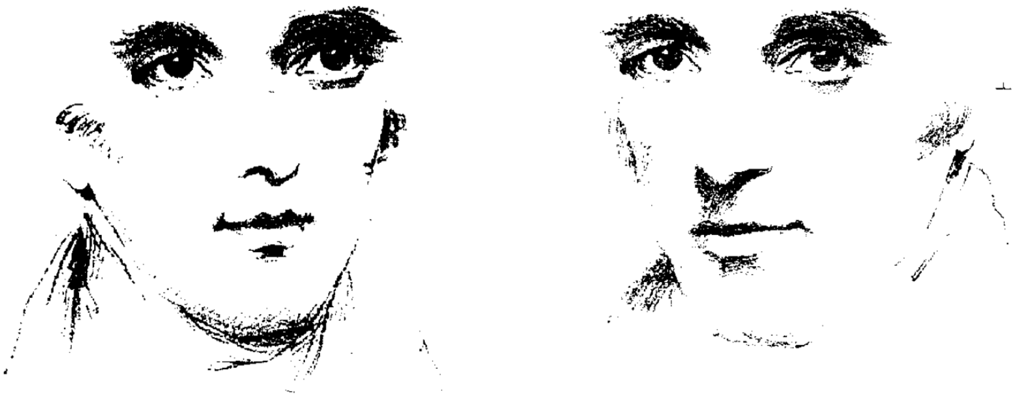


Figure 4.10: Demonstration that head orientation influences the perceived direction of gaze [33]

information to better investigate and analyze facial behaviors of children with ASD. Therefore, the main objective is to find out how the gaze direction and head orientation of individuals would be regulated when an individual with ASD is engaged in a robot-based games. In this regard, the manually labeled eye gaze direction (presented in Section 4.4.2) and automatically tracked head posture of children with ASD (in both speaking and listening contexts) were utilized. The results help us to investigate how often the head orientation and gaze direction of children are following the same directions. Next section presents the captured data and the automatic head direction annotation.

4.5.1 Automated Head Tracking and Attention Level Measurement

Another importatn social cue is head posture and direction of head during conversation. In this regard, this section reports the head orientation of the subjects for both speaking and listening contexts introduced in 4.4.2. To track the direction of head in each video segment, I utilized publicly available IntraFace software [170]. Using IntraFace allowed me to detect and track facial images in sequence of frames. IntraFace also provides 49 facial landmark points as the key component of the face and also estimates head orientation in the given facial image. More details about IntraFace can be found in [170]. Some parameters such as facial occlusion, significant head orientation and illumination variation can reduce the quality and accuracy of head estimation algorithms. I define head orientation as the head posture with respect to a point of interest (i.e. robots position). Figure 4.11 demonstrates a video segment with head orientation and gaze direction annotations.

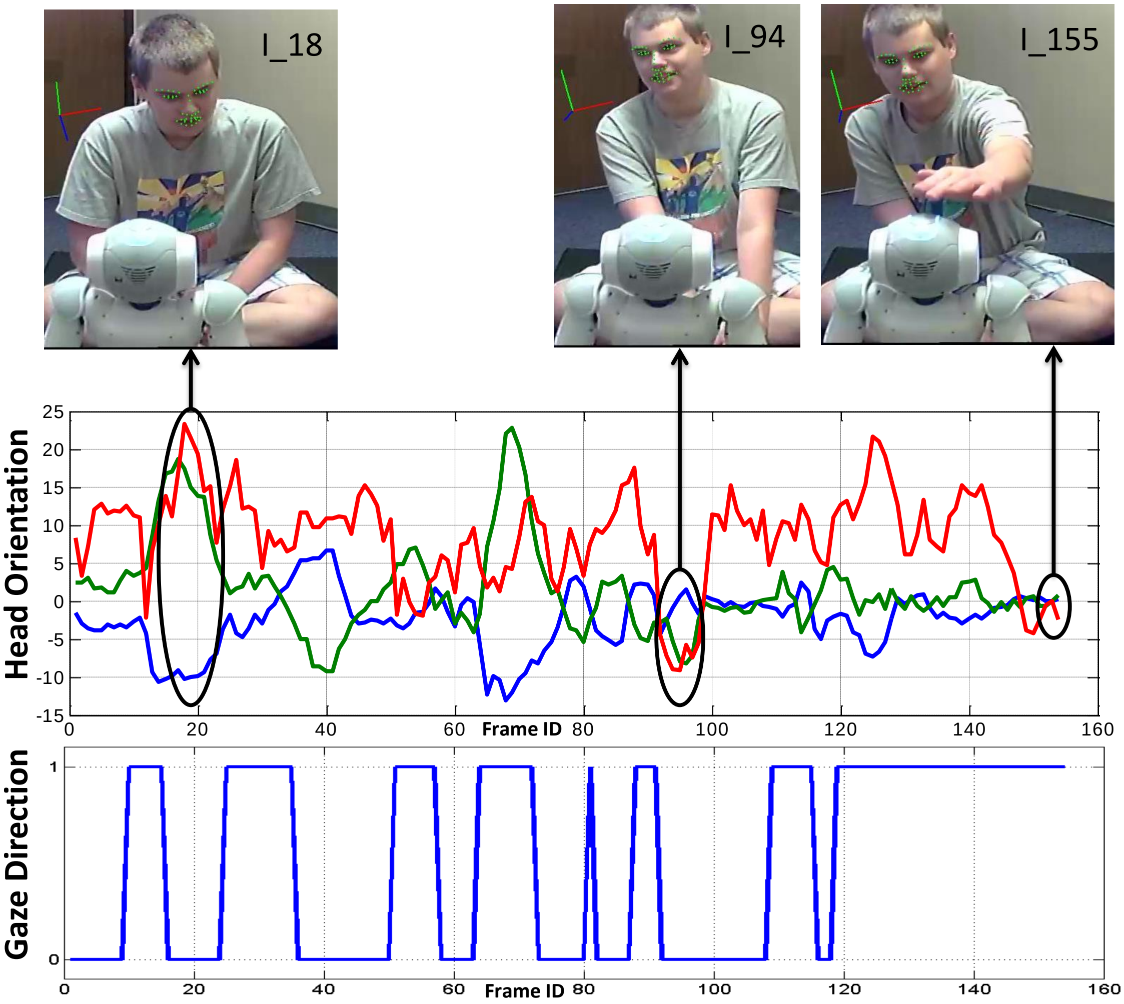


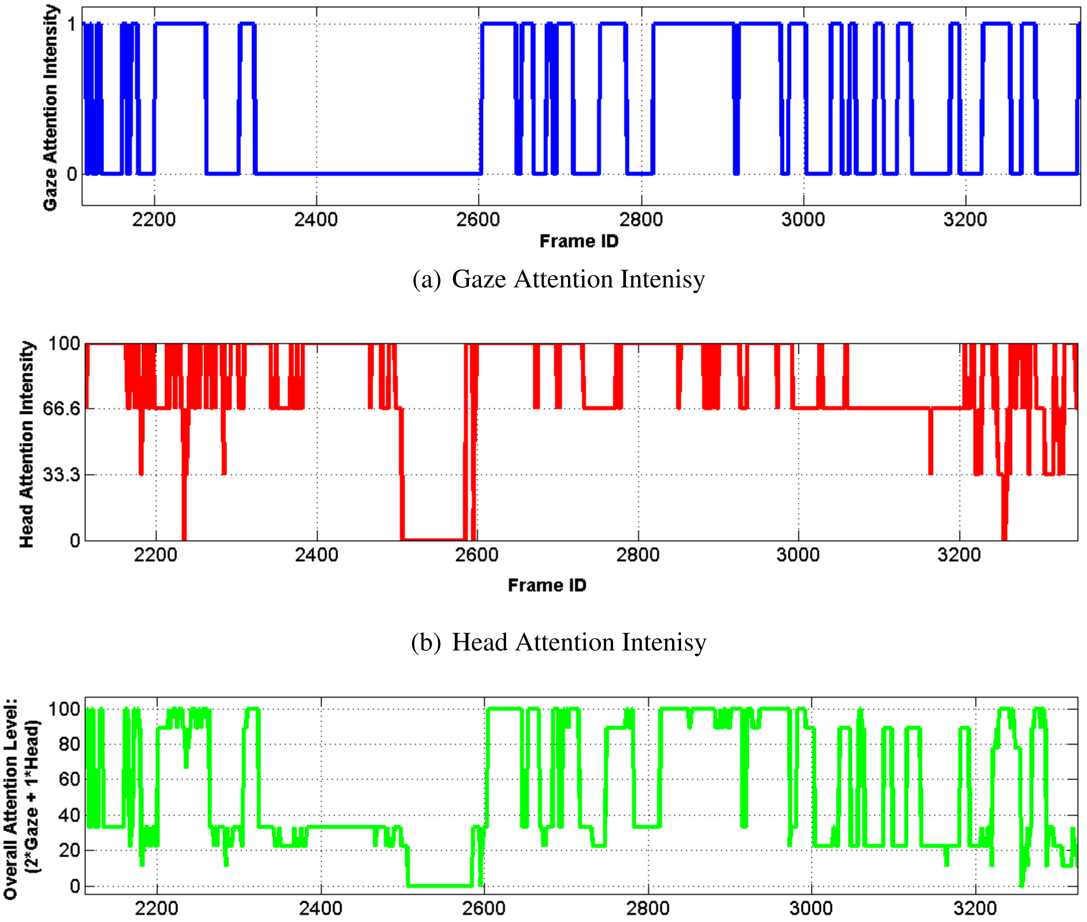
Figure 4.11: Sample segment of video of a child interacting with NAO: Head orientation and gaze annotations are plotted

For this experiment, the video of 11 children with ASD that introduced in section 4.4 were used. Video segments corresponding to the *”child speaking”* and *”child listening”* segments were analyzed using IntraFace. Then, three angles (i.e. α, β, γ) for each facial image were computed. As reported in Table 4.3, the head have more tendency to rotate along z-axis more often (i.e. γ has high mean and STD value). In addition comparing the speaking and listening contexts confirms that the head orientation of children with ASD in listening is much lower than the speaking one.

Table 4.3: Mean and Standard Deviation (STD) of head orientation for children with ASD in conversational context

Social Context of the Game	**Angular Head Orientation**		
	α	β	γ
Speaking (mean)	0.59	-2.73	15.43
Speaking (STD)	5.43	9.29	34.63
Listening (mean)	-0.49	0.36	3.78
Listening (STD)	5.03	9.15	17.03

In order to have a more accurate measurement of the attention level of the subjects, the gaze label and head orientation labels were fused together. In other words, to specify the attention level of children with ASD when they communicate with NAO, the gaze and head orientation information were integrated. As reported in [33] to measure the attention level, gaze information can override head cues; therefore I assigned higher weight to gaze-related attention intensity than the head direction (i.e. eye gaze weight is twice as the head weight when calculating the attention value). In addition to quantify the attention level corresponding the head orientation, I assumed that for each three reported angles, once head is rotated more than a specific degree (e.g. $10\,\text{deg}$) the corresponding head-direction has no attention toward the robot. This concludes that every one of the three angles ((α, β, γ)) has the same effect on measuring the head-based attention measure. Therefore the head-related attention level quantifies by ordinal values withing the range of $[0-100]\%$, with resolution 33%. Figure 4.12 illustrates a corresponding gaze, head attention levels of a segment of (child speaking) segment. Subplot 4.12(c) demonstrates the overall attention level of the video segment which is calculated by fusing the gaze and head orientation data.



(a) Gaze Attention Intenisy

(b) Head Attention Intenisy

(c) Overall Attention Level

Figure 4.12: Attention Level of ASD children in Speaking Context

Based on the described approach, the overall attention level of children toward NAO was calculated. In addition, the between-context attention-level comparison was conducted. The results demonstrate that overall the attention level in speaking context (51.71% (STD = 35.68)) is lower than the listening context (65.24% (STD = 35.44)) and the (unpaired) T-test result (with significance level = 0.001) confirms that. This is another important evidence that confirms that children with ASD pay attention to NAO similar to TD individuals while they interact with another human, and confirms that robots can be employed as assistive agents for children with high-functioning ASD.

Last but not least, to find out how often the eye gaze direction and head orientation are directed to the robot, the percentage of time each of these events happened are reported. Table 4.4 specifies the frequency of: 1) gaze directed toward NAO, 2) direction of child's head toward NAO (i.e. head attention intensity level 100) and 3) both head and gaze are toward the robot. Results indicate that in speaking context, children with ASD tend to avert their gaze and head direction from NAO more frequently, however in listening context both head and gaze are pointing to the robot for more duration. In addition, TD children look at NAO almost twice as much while listening that speaking; however they regulate their head toward NAO similarly in both contexts. Furthermore, in speaking context, 11% of time TD children's gaze and head direction is toward the robot, in contrast to 44% in listening contexts. In ASD group, this amount is 16% and 33%, respectively.

Table 4.4: Statistics about eye gaze and head directions of children with ASD and TD children in HRI system (speaking and listening context)

Social Context	Gaze and Head Direction Toward NAO (Frequency %)		
of the Game	Gaze At NAO	Head pointing to NAO	Both Gaze & Head pointing to NAO
TD (Speaking)	33.50%	71.41%	11.50%
TD(Listening)	64.80%	71.79%	44.24%
ASD (Speaking)	39.80%	41.07%	16.59%
ASD(Listening)	54.64%	65.80%	33.71%

4.6 Protocol B: HRI and Autism Therapeutic Applications

In a multi-disciplinary study, we tested the social responses of seven high-functioning verbal children diagnosed with ASD over the course of six months. A time span of such duration was chosen to study the long term effects of the robot on specific social skills of the participating children. Children in this study were paired with a humanoid robot (NAO) to interact in several social task games. These games were designed to quantitatively measure initial existing social skill deficits and provide feedback for improvement from those initial values. The targeted social skills include:

- **Verbal Communication** through socially-simple scripted conversation and verbal instructions given by the robot to play the games.
- **Joint attention and nonverbal communication** through the robots use of pointing and head movement to direct the childs attention toward targeted objects.
- **Facial expression recognition and mimicry** through games using photographs of individuals showing an emotional expression and the robots feedback on the childs attempt to imitate the expression.

I used a humanoid robot (NAO) in the same room with the similar video-audio capturing setting as presented in Section 4.4.1. Following sections introduce participants and details of robot-based therapy sessions.

4.6.1 Methodology

This section introduces participants and capturing setting for robot-based therapeutic setting.

Participants

Seven (2 female and 5 male) high-functioning children with ASD in the age range of 6-15 years old (Mean=10.14; STD = 2.27) participated in this study. All children with ASD were recruited from different autism treatment organizations from the state of Colorado (USA), including local autism schools and families associated with JFK Partners . All of the participants had a high level of verbal communication capabilities. All parents of our participants signed an informed consent form and IRB approval was obtained for our study.

4.6.2 Social Robot-based Games:

To design robot-based games for children with ASD, the goal was to keep every game as exciting and socially interesting as possible. Therefore, each session was divided into five distinct games, each of which was focused on one specific aspect of social responses. To keep the energy level of the child high in every session we asked the participants to have a short break after finishing each game of the session. During each game the participants had the opportunity to illustrate and practice a specific set of social skills (e.g. verbal capabilities, gaze attention, joint attention, facial expression recognition and imitation, and body gesture imitation skills). Throughout the game, NAO referred to the participant by his/her name and also used simple words to instruct and command each participant in as clear and simple manner possible. Sample images of different subjects interacting with NAO in various games are illustrated in Figure 3. Bellow the content of each game has been discussed in more details.

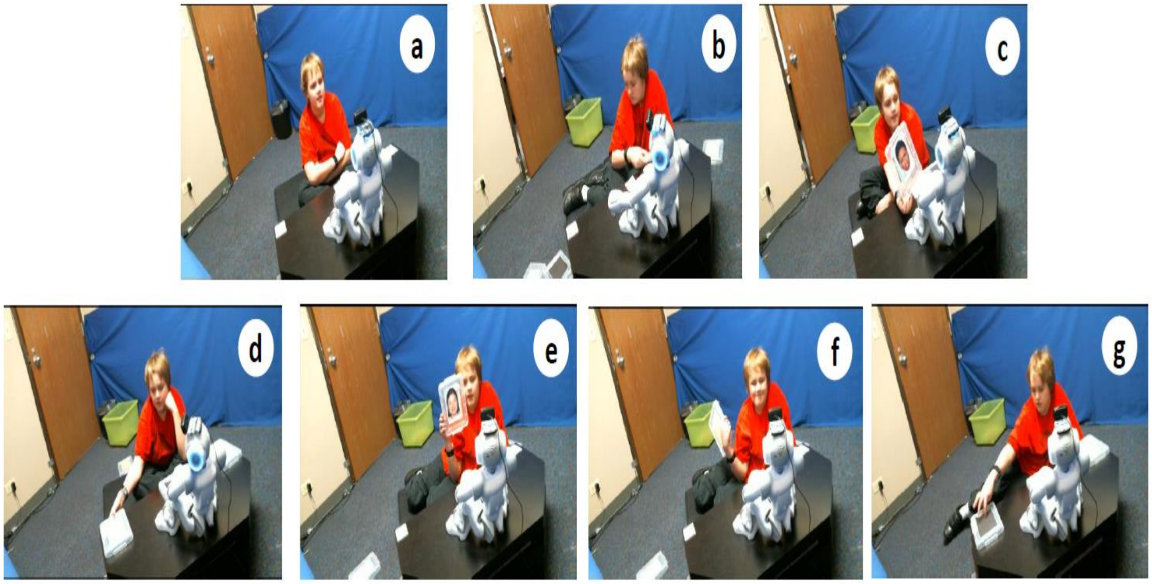


Figure 4.13: Sample images of five social-behavioral games (Game 1: (a) Verbal interaction- Game 2: (b) NAOs pointing, (c) Expression recognition- Game 3: (d) Joint attention- Game 4: (e) Expression recognition, (f): Expression imitation- Game 5: (g) Childs pointing)

Game 1: Verbal Communication with NAO: The goal of this game was to build a friendly relationship between the child and NAO at the beginning of each session. Hence, the participants were asked to listen to NAOs instructions and respond to five questions related to his/her full name, favorite colors, number of siblings, age, and time the child goes to bed. Before starting the games, we provided a questionnaire from the parents to get the response to all these questions and document how well or consistently every child responded to the questions (the results related to Game 1 are represented by G1 in the rest of the chapter).

Game 2: Following NAOs Direction of Pointing and Recognizing Facial Expression: The goal of this game was two-fold. First, we intended to evaluate how accurately the participant could follow the direction that NAO was pointing toward and pick up the objects requested by NAO pointing at them (G2.1). Second, we examined how each individual could recognize various facial expressions (by providing list of possible facial expressions) and what categories of facial expressions each individual had difficulty recognizing (G2.2). In this game, NAO started by introducing the child how to play. NAO pointed to a box in the room, which contained a picture of an expressive face inside. NAO then asked the child to bring the box and show NAO the image inside the box. Thereafter NAO asked the child What facial expression is on the image? In this segment of the game, NAO assisted the child by providing multiple choices (e.g. happy, sad, angry, and neutral) for the child to label the facial expression shown by the individual on the photo.

Game 3: Joint Attention Skill: For this game (G3), we focused on the joint attention skill (i.e. two conversational partners looking at the same point or object by means of gaze and/or gesture [94]of participants with ASD. To investigate how children can regulate their point of attention toward an object of interest, NAO asked the child to bring the box NAO was looking at. Once NAO turned its head toward a box the child needed to tell NAO which box it is focused on. This was repeated five times for different boxes in the room.

Game 4: Facial Expression Recognition and Imitation Skills: We placed a few boxes inside the room that contained expressive facial images. During this game NAO pointed to one box at a time. The child was then instructed to recognize the facial expression shown in the photo inside the box (G4.1). In contrast to G2.2 we did not provide a list of facial expression choices. Finally, NAO instructed the child to carefully look at the image and try to imitate the expression on the child's own face (G4.2). The same procedure was repeated for five different boxes.

Game 5: Child Pointing Skills: For this part of the study (G5), NAO described a box in the room and the child was instructed to point toward that box. This segment of the game aimed at measuring how well the participant could understand the verbal commands of NAO and the child's ability to direct the attention of the robot toward the described object. Therefore, NAO described five different boxes and the child was instructed to point at every one of them.

After each game the child was given the option to take a break for 5 minutes if desired. This break allowed the child to maintain concentration for the next segment of the game so as not to exhaust the child's attention span. The entire session for each individual took around 45-60 minutes. In every session of the game, a graduate researcher observed and annotated the responses of the child and calculated the accuracy of responses in all five games within the range of [0-100]. This accuracy index is a quantitative measure that we used to assess the participant's skill development for every human-robot interaction session.

4.6.3 Experimental Procedure and Results

Our longitudinal multi-session study was divided into three phases: 1-Assessment, 2-Intervention and 3- Follow-Up. A description of each phase is given below: Assessment: The first three sessions of the child-robot interaction were used for assessments. In these sessions, we evaluated the current knowledge and accuracy response rate of every child by having them interact with NAO in the five aforementioned games without any feedback to the child. The captured data from these child-robot interaction sessions served as baseline data to examine the developmental skills of each participant (e.g. social interaction, cognition, gaze direction and pointing regulation, facial expression recognition and mimicking skills). For each participant, if the accuracy of any of the games was less than the 80% threshold we conduct the intervention sessions for that game. Otherwise no assisting feedback was given for that specific game.

Intervention/Treatment/Experiment: Based on the baseline results of each individual, we evaluated whether the child needed intervention sessions or not. If the child had a low level of performance (i.e. lower than 80% accuracy) in one or more game segment(s), we tailored the session so that NAO would give intervention to the child in that game (i.e. one game at a time) and kept all the remaining games without feedback. We repeated the same procedure for the selected game until the performance reached 80% or more, and then moved to the next game which needed intervention. Having intervention for only one game at a time allowed us to observe if the child's skill accuracy for any other game was also affected. For the games where intervention was needed, NAO provided verbal feedback and further explanation that could help the child better understand and respond to the questions or instructions of the game. Follow-up: In the final phase of the study, we conducted two follow-up sessions with human-human interaction without any feedback to evaluate how well the child could sustain, generalize, and apply the learned concepts when NAO is substituted by a human. Throughout these sessions, the researcher who operated NAO followed the same procedure and verbal commands as given with the robot in previous stages of the game.

Subject-based Therapeutic Results

We designed and conducted a longitudinal robot-based therapeutic setting for 7 children with ASD. Each individual participated in our study for about 6 months and interacted between 6 to 17 sessions with NAO. The first session of interaction was assigned to just familiarize each child with the robot and give him/her the chance to communicate and play with the robot. Then we started the baseline sessions to assess the initial social knowledge of every child. The assessment results indicated that 6 of the 7 children with ASD have some difficulties responding to one or more of the games (i.e. accuracy response lower than 80%). Our baseline data clarified that all the subjects responded to G3 and G5 with a high accuracy (i.e. following NAO's gaze and head orientation and also pointing to the described object, respectively). However some of the subjects had difficulty responding to G1, G2, and/or G4. Bellow we discuss the graphical and statistical results of subjects during different phases of our study and illustrate the effectiveness of using robots to assist individuals with ASD and improve their socio-behavioral deficits.

Figure 4 demonstrates the responses of one of the subjects (i.e. SN024) for two games which needed intervention (i.e. G1 and G2) during three phases of our investigations. The average baseline accuracy for G1 and G2 were 33.33% and 0% respectively. After conducting 12 sessions of intervention by NAO, the accuracy of both games (G1 and G2) reached 80%. In the follow-up human-human session we observed that the responses of SN024 stayed in a similar accuracy range. The graphical representations of the behavioral responses of all seven participants of our study are reported in the Appendix A.2.

To statistically evaluate how the robot-based intervention can improve the social and behavioral responses of participants with ASD, we conducted t-test analyses (significance level 0.05). Since only six of the participants needed intervention for one or two games, we report the t-test results of 5 out of 7 children who went through the intervention for G1. We calculated the average of three baseline sessions and the follow up sessions of each partici-

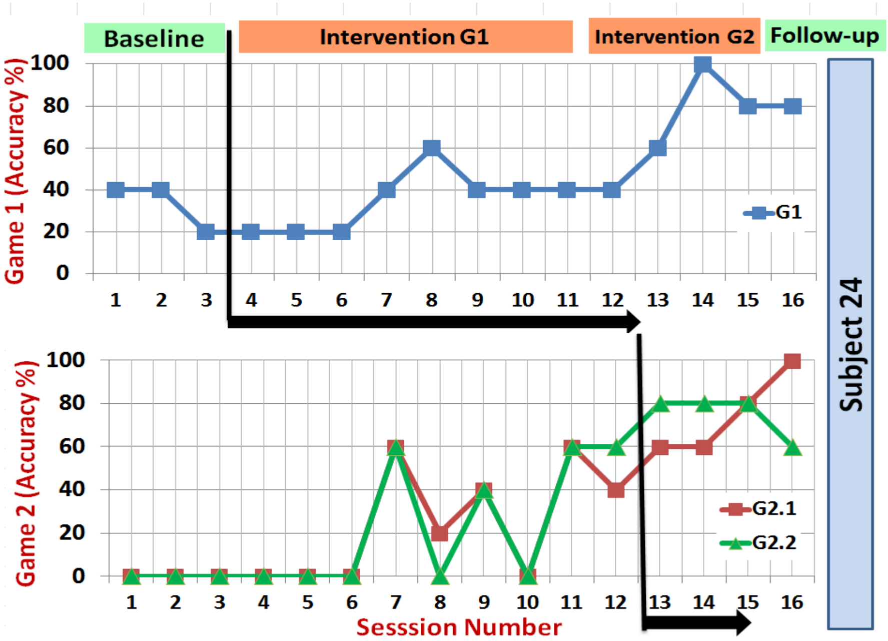


Figure 4.14: The accuracy of behavioral responses of SN024 for Game1 and Game2: the vertical line demonstrates the starting session of each intervention and the horizontal arrows specify the duration of the intervention for each game.

pant and compared them with the last intervention session for G1. The paired t-test results for all the subjects confirmed that the responses of children with ASD after conducting the robot-based treatment (or intervention) were significantly improved compared to the baseline data (t(4)= 3.474, p=0.013). In addition, analyzing the intervention and follow-up results for G1 indicates that the children with ASD are capable of generalizing the learned skills by robot treatment and apply them during the human interaction. Statistically, for G1 there was no significant difference between the intervention and follow-up sessions for ASD individuals.

Table 4.5 reports the overall performance (i.e. average and standard deviation of accuracy of responses for five social-behavioral games) of all seven children with ASD in three phases of our study (i.e. Baseline, Intervention, and Follow-up). Comparing the baseline and intervention results confirms that for a majority of the social games, using the robot-based intervention can improve the desired social and behavioral skills of the children for all five games. Interestingly, the results of human-human follow up sessions indicate that after learning the social skills by the robot, children are able to transfer the learned skills to the humans interaction and respond to the human with an even higher accuracy than the robot interaction.

Table 4.5: Behavioral response accuracy: Average (STD) of all seven children in three phases.

Game ID	**Overall Response Accuracy: Average (STD)**		
	Assessment	**Intervention**	**Follow-up**
Game 1	79.05(27.87)	81.95 (30.05)	98.33 (25.07)
Game 2.1	80.00(38.35)	80.49 (37.99)	100.00(25.07)
Game 2.2	61.90(36.23)	75.61 (38.68)	96.67(15.74)
Game 3	82.86(37.57)	–	96.67(37.80)
Game 4.1	66.67(36.23)	72.68 (44.57)	100.00(9.76)
Game 4.2	65.71(33.94)	60.98(38.82)	80.00(33.52)
Game 5	92.38(18.44)	–	100.00(0.00)

Results reported in Table 4.5 demonstrate the effectiveness of robot-based sessions for teaching and improving facial expression recognition (FER) skills by children with ASD. The baseline data for G2.2 (i.e. FER by providing list of facial expression), the low FER rate (61.90%) with high variability (STD=36.28) demonstrate that overall, participants had difficulty recognizing expressions. The intervention results, confirm that a robot can teach children recognizing facial expressions reliably (higher accuracy). Statistical analysis of G2.2 reveals that there is significant improvement in facial expression recognition using the robot-based intervention ($t(1) = 7, p = 0.045$). The results of follow-up validation phase indicate that children could recognize the basic facial expressions with a high accuracy (96.67%) and a very limited variability (*STD = 15.74*) and statistically there is no significant difference between intervention and follow-up sessions (*p-value =0.25*). We can make similar discussion for Game 4.1 as well. These results conclude that robot-based feedback and intervention with a customized protocol for each individual can improve the learning capabilities and social skills of children with ASD.

4.6.4 Discussion: Robot-based Therapy

Autism spectrum disorder is referred to a wide range of deficits that vary from the lack of social and conversational skills and may cause symptoms, such as aggression, depression and anxiety. Individuals with autism can benefit from the treatments when not only their core deficits are identified as soon as possible, but also through participation in individualized intervention sessions. Several studies published over the last few decades have demonstrated that many children with autism who receive intensive ABA interventions make substantial growth in their learning, adaptive skills, and behaviors [159]. In our study we designed an individualized robot-based intervention sessions based on ABA to systematically observe, analyze and measure the behavioral and social responses of individuals in longitudinal multi-session games. The objective of our study is to conduct games for each participant to teach behavioral skills to the children with autism and improve their socio-behavioral skills. We began our treatment procedure by conducting a comprehensive assessment of the childrens developmental status. Results from this evaluation helped us to measure the social and behavioral responses of each participant as the baseline data. In addition, such assessment allowed us to develop a robot-based treatment program that was tailored to each childs strengths and needs. As it is true for the majority of individuals with neurodevelopmental disorder [113], for children and adolescents with ASD training and interventions for improving social and behavioral skills are crucial to support independence and functionality in the society [47, 135]. Therefore, we designed and conducted a set of therapeutic games were the robot, NAO, responded and commanded the child to help the child if required during each game. To evaluate how NAO is capable of teaching new social and behavioral skills to each participant, (after finishing the intervention sessions) we followed up our investigation with a human-human interaction. The human-human interactions used the same games and session structure as the robot-child interactions. The follow-up sessions allowed us to investigate how each child could generalize the new learned skills to human interaction. Our graphical results depicted in the Appendix, demonstrate that all of the children showed considerable understanding and appropriate responses toward following the eye gaze and head orientation of NAO (G3) or pointing to an object described (G5). However, the majority of them lacked the ability to consistently respond to verbal questions (G1) and also had deficiencies in recognizing and imitating facial expressions (G2 G4). By comparing the results reported in Table 1, we can confirm that overall the robotic treatment can be a viable approach for helping individuals, once we know exactly what the core deficits of each individual are and personalize a set of games to help an individual to practice them. Based on the multiple robot interaction sessions and after analysis of the recorded videos of the study, we noted several observations that may be helpful for those who would like to design ASD intervention therapies.

- Some of the participants were very engaged with NAO and they often talked and communicated with the robot well, but they would become frustrated if NAO took too long to respond or repeated an instruction after the child had already completed the task.

- In terms of facial expression recognition (G2.2 and G4.1), most of the subjects were highly capable (except SN024), and occasionally one subject (SN019) had difficulty recognizing neutral as sad and imitating a sad face when shown a neutral face. Another subject (SN020) mistaken sad for angry on multiple occasions, even when Nao provided options for the expressions in G2.

- For facial expression imitation (G4.1) a few subjects (e.g. SN019 and SN023) had trouble maintaining an expression on their face for an extended time (about 5 seconds), and oftentimes flipping between expressions very quickly. For instance SN019s face jumped between happy and sad faces while trying to mimic happy, or throwing in bits of surprise (raised eyebrows specifically) when trying to maintain a neutral or sad face. Some of the subjects (e.g. SN020 and SN24) were almost entirely inaccurate in terms of mimicry. They made a happy face for every expression presented until the last few sessions, where NAO explained how to build the expression during the intervention sessions.

- Coding behavioral and social responses of children with ASD could be a very challenging task. For instance, some of the subjects (e.g. SN019, SN024, and SN025) for a few sessions moved around a lot or even hit NAO repeatedly and moved the camera around. During a few sessions they got very close to the camera sometimes on top of NAOs head, looked down a lot or obstructed the view of their face making it difficult to monitor, observe and annotate their facial behaviors.

Pertaining on observation and experimental results of our study, it would be worthwhile investigating the following points and questions in future studies:

Question: Do children with ASD, ever moved around or hit the robot, either as a sign of frustration (on difficulty levels) of the games, or just because the interlocutor partner (i.e. robotic or human) responded slowly or not quickly enough to the children?

One approach can be to design a robot-based game with different difficulty levels to even further tailoring or personalization of the social skill games, and evaluate if the child would show similar negative traits or behaviors during the interactions.

Question: How children with ASD would react differently in robot-based intervention versus conventional human-based setting?

In our proposed protocol, we started the experiment by conducting the human-robot baseline assessment and after intervention sessions, the results of human-human follow up sessions were reported and analyzed with previous stages. For future research directions, we would suggest to begin the experiment with the human-human baseline and then human-robot baseline sessions to investigate how the social responses of children can be improved in between settings. Moreover, this allow us to investigate in which behavioral and social settings children with ASD may gain more benefits from robot-based therapeutic protocols.

In addition, it is worth mentioning that to having a control ASD group for similar human-human therapeutic setting help researchers to better evaluate the efficiency of robot-based therapeutic ssettings. Recruiting ASD children who are only involved in longitudinal human-based therapy session and comparing them with the ASD group that are involved in the same treatment procedure, but robot-based, is highly recommended. The control group would help researchers to find out how and in what sense human-human or human-robot therapy sessions are more effective for different social and behavioral treatments. This will open new trend for better fusing robot-based protocols into conventional human-based therapy sessions.

Question: Do individuals with ASD have similar behavioral and social responses when a robot has a dynamic (emotional) face? In our experiment we used NAO, a humanoid robot with a static face. However, would a facially expressive robot (e.g. Zeno [79]), improve and accelerate the social learning skills of the children?

The main objective of using robots as a therapeutic agent is to train individuals with ASD for having engaging human-human interactions. Hence, it is important to investigate what social cues (e.g. facial expression, eye gaze, head orientation, etc.) made it complex and difficult for ASD individuals to an engaging sociable interactions with other humans. In addition, having a human baseline in the beginning of the experiment would be very valuable information for therapeutic and robotic perspectives to better certify the impacts of robot interventions.

4.7 Conclusion

Studies in autism research demonstrate that individuals with ASD tend to have more engagingly interactions with robots than humans. This chapter reports the result of our longitudinal therapeutic robot-based protocol which was designed to assist the children with autism and improve their social responses. The results demonstrate that in overall, robots can be viable approach for teaching some basic social skills to individuals with ASD. As the gaze regulation and head orientation play important roles in conversational context, they have been investigated as well. First, the gaze responses of TD and ASD groups were compared in speaking and listening contexts. Calculated statistics confirm that overall the gaze direction of both TD and ASD individuals follows the same pattern, although the ASD group tend to have less amount of time fixing their gaze toward the robot. In addition, using VMM algorithm confirms that automated algorithm are powerful technique to discriminate between the gaze responses of TD and ASD groups. Lastly, the head orientation of ASD children studied and the gaze and head orientation information were used to measure the attention level of individuals in HRI system. The results confirm that when children with ASD interact with NAO, they show similar characteristic as TD individuals when they interact with another interlocutor partner. The facial behavioral analysis of ASD children in this section concludes that although there are some differences between social responses of TD and high-functioning ASD children in HHI, the ASD children interact with robots very similar to TD peers in HRI system.

Chapter 5

Discussion, Conclusion and Future Work

In this dissertation, I investigated human's facial behavior computing in two important HMI settings: 1) Computer-based, and 2) Robot-based. The main objective was to study and model spontaneous facial cues (e.g. facial expressions, gaze and head orientation) in real-world HMI systems and explore how machine learning techniques can automatically compute and analyze user's responses.

This dissertation is the first work that introduces a publicly available database to investigate the dynamics of facial muscle movements in spontaneous facial expressions. In addition, it provides a baseline protocol and benchmark for automated action unit intensity measurement in future research. Furthermore, it compares and investigates AU temporal dynamics of posed and spontaneous facial expressions using statistical computational techniques and automated algorithms.

As the second objective, this dissertation investigates facial behaviors of children and adolescents in a real-world human-robot interaction setting. We first introduced a robot-based assistive protocol for interacting, understanding and improving social and behavioral responses of children with high-functioning autism. Then, we developed and deployed a robot-based intervention protocol for autism therapeutic applications and modeled facial behaviors of children in a real-world HRI system. The following sections summarize the achievements, challenges and future research directions of this dissertation.

5.1 Summary of Dissertation Achievements

5.1.1 DISFA: a spontaneous FACS-annotated database with AU dynamics

Because of the importance of facial expressions in face-to-face communication and lack of spontaneous facial expression dataset, we captured and annotated spontaneous facial expressions in HCI setting and named it DISFA database. DISFA describes facial expressions in terms of AUs and provides the intensity-level of 12 AUs for about 130,000 frames. We believe DISFA can allow researchers to study the dynamics and temporal patterns of facial muscle movements comprehensively. DISFA also provides a common ground for formulating and investigating new approaches for facial image analysis. Since the time DISFA released (November 2013) till this day, over 80 national and international research centers and universities have requested to download DISFA.

5.1.2 Automated spontaneous facial expression computing

We designed and implemented automated systems for computing the facial expressions and their dynamics in a single image or sequence of facial images. Although we are not the first group to present a fully automated facial expression recognition system, this work is among those studies, which comprehensively analyzed and modeled spontaneous facial AU dynamics (e.g. recognize presence, occurrence and intensities of AUs). In our experiments we used SVM (and HMM-based) classifiers along with the Gabor feature extraction and manifold learning techniques to automatically recognize AU dynamics. According to the reported results in section 3.3, the proposed algorithm is capable of recognizing the dynamics of AUs with high reliability (Overall reliability $ICC = 0.77$).

5.1.3 Comparing spontaneous vs posed facial expressions

There are several available studies about posed facial expressions and designing automated system for recognizing prototypic expressions or AUs in non-genuine setting. However, there are only few works, which compare posed and spontaneous facial expressions and to the best of our knowledge there are no available work that compares the dynamics and temporal patterns of posed and spontaneous facial AUs.

Therefore, as a part of this dissertation, I developed and used a software to capture posed facial actions of adults (a sub-set of DISFA participants were rejoined for DISFA+ recording) and captured and annotated their 12 AU dynamics. I statistically compared the temporal and co-occurrence characteristics and used t-test to find the differences of posed and spontaneous facial expressions. In addition, I designed an automated system to represent, model, and classify the posed and spontaneous facial expressions.

5.1.4 Autism and robot-based therapy agent

Another objective of this dissertation was to design a robot-based protocol in real-world therapeutic applications. Since children with autism show interest toward technology, we designed and conducted a set of robot-based games to study and foster the socio-behavioral responses of children diagnosed with high-functioning ASD. The behavioral responses of participants during different phases of this study (i.e. baseline, intervention and follow-up) reveal that overall, a robot-based therapy setting can be a viable approach for helping individuals with autism.

5.1.5 Facial behavior analyses of children with ASD in HRI

Considering the fact that gaze direction and head orientation patterns illustrate the emotionally interest or engagement intensity of an individual in a dyadic social communication, we modeled the facial behaviors of children with ASD in HRI. The goal was to examine and model gaze responses of children with ASD and Typically Developing (TD) children in conversational context with NAO. In addition, in the second phase of this study, we developed an automated system to track the head orientation of subjects during robot interaction. Integrating both the eye gaze direction and head motions allowed us to model the attention level of children during human-robot interaction. The results confirm that ASD children similar to their TD peers have higher attention toward the robot when they are listening to NAO. In other words, their gaze and head direction are pointing toward the robot for 33% while listening than 16% when they are speaking (reported in Section 4.5.1).

5.2 Discussion and future work

While we have reported the results and made a number of improvements to the current state-of-the-art automatic facial expression recognition algorithm and HRI systems for autism research, a number of problems and challenges still remain unsolved and need to be addressed in future research.

5.2.1 Facial expression analysis: challenges and outlook

Spontaneous facial expression recognition is an important research area that has recently grasped the attention of the research community. As mentioned earlier, the lack of manually coded spontaneous facial behavior database has hampered research in automated facial expression recognition. DISFA, which is introduced in this dissertation and released publicly can be applied to study the dynamics, intensity, and spatiotemporal structure of spontaneous FACS action units. Besides the data collection, measurement and coding of spontaneous facial expressions and transforming the knowledge from posed to the spontaneous domain are still challenging. In the following part, we introduce some of the intrinsic challenges one might encounter in studying facial expressions, along with some suggestions for future research.

- Co-occurring action units may or may not present a challenge. Eye movement and blinking make the face annotation harder when they co-occur with the AUs that describe the eyes actions. For instance, eye movement can influence the detection and intensity measurement of AU5. In other words, in AU5 with maximum intensity (i.e. AU5(E)), the upper eyelid is pulled up maximally, revealing a maximum amount of sclera above the iris. In this case an eye movement can influence the upper lids muscles and AU5. Considering that, to improve the quality of automated facial muscle movement or emotion recognition, one may consider other AUs or action description (e.g. head movement, eye movement).

- Another issue arises from the fact that different types of features may be more or less revealing of different AU. Some AU create bulges or pouches (e.g. AU14, AU15, AU24); others create lines, wrinkles, or furrows (e.g. AU12, AU9). And others reveal new features (e.g., teeth in AU 25) or occlude existing features (e.g., closed eyes in AU 43 or 45). Some AU may produce multiple changes. Obviously, a single facial feature representation cannot describe all these appearance changes appropriately. In our work, we utilized facial image representations, like LBPH, HOG and Gabor features. Discovering a set of features that has both powerful representing and discriminating capabilities can be another interesting research topic in the field of (spontaneous) facial expression recognition. With DISFA, one could identify optimal features for each AU, which could increase efficiency of AU measurement.

- In DISFA pilot coding 12 AUs were coded. Other researchers may harvest additional AUs. We would encourage that effort and ask that they vet their coding and make it available to the community.

- The frame rate in DISFA was set to 20 fps. The frame rate was chosen to accommodate the hardware synchronized acquisition with resolution of 1024×768 pixel for both left and right images. While this necessitated a slower frame rate than is usual, motion artefact or discontinuity was absent. This frame rate may be considers as a not sufficient frame rate for analyzing the very subtle facial expression analysis such as micro expression analysis. Therefore, one might need to obtain videos with very high frame rates for those types of applications.

- We realized that some action units occurred less frequently than others (see Table 3.3). Hence, if there are not enough number of samples to train a classifier then we will have a less reliable system. In this regard, designing a system that can be trained with a small number of training samples is desirable.

- One important application of automated AU dynamics computing, is categorizing posed from spontaneous facial actions. In section 3.7.2, we have shown that the velocity of some muscle movement in rising and decaying duration are significantly different. This concludes that considering the temporal characteristic of AUs can be a helpful information to design a more effective automated FER system.

- In last two decades several automated algorithm and database were released to categorize posed facial actions. However there are very few studies for spontaneous facial expression. One critical solution for better computing spontaneous facial expressions is transferring the learned knowledge from the posed domain to the spontaneous domain.

5.3 Autism and HRI systems: problems and future directions

This dissertation investigate using HRI systems in robot-based therapeutic application. We programmed and remotely controlled a robot to interact with set of children with hight-functioning autism for a longitudinal study. During over a year of interaction with high functioning children with autism, we can indicate and report the following observations, that may be helpful for other researcher in the field:

- In general, children with ASD having appropriate patterns of gaze-direction and attention regulation toward NAO during robot-based games. However, some of them became frustrated or bored easily when the robot delayed responding to their requests, or once NAO repeated the questions multiple times. These aspects need to be considered if one is interested to use robots an a therapy agent.
- As NAO had an static face, in order to measure the capabilities of children in recognizing and imitating expressions, we used some expressive photos. Our results indicate that majority of children with ASD have difficulty imitating facial prototypic basic expressions. A robot with an expressive face might be a good choice for interacting with children.
- Participants had a diverse range of reactions to the same games and some children moved around the room and hit the robot, or talk to the robot without paying attention to the robot's commands. It is important to conduct some fundamental studies to

understand what kinds of games or social-therapy sessions would best serve autistic individuals.

Considering all these facts, I aim to investigate how we can integrate the robot-based assistive agent to the existing therapeutic settings. Here are few questions that addressing them can shed light on this research field:

- Whether individuals with autism hold the same gaze engagement patterns when the robot has a dynamic emotive face with various facial expressions or not?
- Which types of individuals with ASD can best take advantage of using technologies and robots?
- How well an individual can translate the skills learned with a robot to a life-long human-human interaction?

Considering the existing technologies and all the aforementioned challenges and limitations for using robots in therapeutic settings, we emphasize that more comprehensive studies need to be accomplished. These studies can raise the awareness about what features in robots attract individuals with ASD most, and how therapists, families, and the community can take advantage of the technology to better understand and interact with individuals with autism.

Bibliography

[1] ”http://affect.media.mit.edu/areas.php”.

[2] ”http://www.aldebaran.com/en/”.

[3] ”http://pthumbnails.5min.com/10342942/517147095_c_o.jpg”.

[4] ”http://pthumbnails.5min.com/10342942/517147095_c_o.jpg”.

[5] ”http://i.ytimg.com/vi/supsc47NivI/maxresdefault.jpg”.

[6] ”http://www.toyota-itc.com/activities/img/5img5_b.png”.

[7] ”http://pthumbnails.5min.com/10342942/517147095_c_o.jpg”.

[8] ”http://www.scienceofpeople.com/2013/09/guide-reading-microexpressions/”.

[9] ”http://www.paulekman.com/micro-expressions/”.

[10] ”http://www.engr.du.edu/mmahoor/DISFA.htm”.

[11] ”http://www.cdc.gov/ncbddd/autism/data.html”.

[12] ”`http://www.autismspeaks.org/what-autism/diagnosis/dsm-5-diagnostic-criteria`”.

[13] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *Computer vision-eccv 2004*, pages 469–481. Springer, 2004.

[14] Neşe Alyüz, Berk Gökberk, Hamdi Dibeklioğlu, Arman Savran, Albert Ali Salah, Lale Akarun, and Bülent Sankur. 3d face recognition benchmarks on the bosphorus database with focus on facial expressions. In *Biometrics and Identity Management*, pages 57–66. Springer, 2008.

[15] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[16] Michael Argyle and Mark Cook. Gaze and mutual gaze. 1976.

[17] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. The painful face–pain expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788–1796, 2009.

[18] Pashiera Barkhuysen, Emiel Krahmer, and Marc Swerts. The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America*, 123(1):354–365, 2008.

[19] Lisa Feldman Barrett. Was darwin wrong about emotional expressions? *Current Directions in Psychological Science*, 20(6):400–406, 2011.

[20] Marian Stewart Bartlett, Gwen Littlewort, Ian Fasel, and Javier R Movellan. Real time face detection and facial expression recognition: Development and applica-

tions to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, pages 53–53. IEEE, 2003.

[21] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 223–230. IEEE, 2006.

[22] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573. IEEE, 2005.

[23] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573. IEEE, 2005.

[24] Marian Stewart Bartlett, Gwen C Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, and Javier R Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[25] Geoffrey W Beattie. Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology*, 17(1):7–15, 1978.

[26] Ron Begleiter, Ran El-Yaniv, and Golan Yona. On prediction using variable order markov models. *J. Artif. Intell. Res.(JAIR)*, 22:385–421, 2004.

[27] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[28] Vinay Bettadapura. Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*, 2012.

[29] Steven M Boker, Jeffrey F Cohn, Barry-John Theobald, Iain Matthews, Michael Mangini, Jeffrey R Spies, Zara Ambadar, and Timothy R Brick. Something in the way we move: Motion dynamics, not perceived sex, influence head movements in conversation. *Journal of Experimental Psychology: Human Perception and Performance*, 37(3):874, 2011.

[30] Fabrice Bourel, Claude C Chibelushi, and Adrian A Low. Robust facial feature tracking. In *BMVC*, pages 1–10, 2000.

[31] Fabrice Bourel, Claude C Chibelushi, and Adrian A Low. Recognition of facial expressions in the presence of occlusion. In *BMVC*, pages 1–10. Citeseer, 2001.

[32] G. Bradski. *Dr. Dobb's Journal of Software Tools*.

[33] Vicki Bruce and Andy Young. *In the eye of the beholder: The science of face perception.* Oxford University Press, 1998.

[34] Ross Buck, Robert E Miller, and William F Caul. Sex, personality, and physiological variables in the communication of affect via facial expression. *Journal of personality and social psychology*, 30(4):587, 1974.

[35] Lawrence Cayton. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, pages 1–17, 2005.

[36] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[37] Claude C Chibelushi and Fabrice Bourel. Facial expression recognition: A brief tutorial overview. *CVonline: On-Line Compendium of Computer Vision*, 9, 2003.

[38] John G Cleary and William J Teahan. Unbounded length contexts for ppm. *The Computer Journal*, 40(2 and 3):67–75, 1997.

[39] Marvin G Cline. The perception of where a person is looking. *The American journal of psychology*, pages 41–50, 1967.

[40] Ira Cohen, Nicu Sebe, Ashutosh Garg, Lawrence S Chen, and Thomas S Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1):160–187, 2003.

[41] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[42] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Psychology Press, 1988.

[43] Jeffrey F Cohn. Advances in behavioral science using automated facial image analysis and synthesis [social sciences]. *Signal Processing Magazine, IEEE*, 27(6):128–133, 2010.

[44] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, pages 203–221, 2007.

[45] Jeffrey F Cohn and Karen L Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2(02):121–132, 2004.

[46] Timothy F Cootes, Gareth J Edwards, Christopher J Taylor, et al. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[47] Eric Courchesne, CM Karns, HR Davis, R Ziccardi, RA Carper, ZD Tigue, HJ Chisum, P Moses, K Pierce, C Lord, et al. Unusual brain growth patterns in early life in patients with autistic disorder an mri study. *Neurology*, 57(2):245–254, 2001.

[48] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[49] Charles Darwin. *The expression of the emotions in man and animals*. Oxford University Press, 1998.

[50] Kerstin Dautenhahn. Design issues on interactive environments for children with autism. In *In: Procs of ICDVRAT 2000, the 3rd Int Conf on Disability, Virtual Reality and Associated Technologies*. University of Reading, 2000.

[51] Kerstin Dautenhahn and Iain Werry. Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmatics & Cognition*, 12(1):1–35, 2004.

[52] Fernando De la Torre, Joan Campoy, Zara Ambadar, and Jeff F Conn. Temporal segmentation of facial behavior. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[53] Fernando De la Torre and Jeffrey F Cohn. Facial expression analysis. In *Visual Analysis of Humans*, pages 377–409. Springer, 2011.

[54] Disabilities Monitoring Network Surveillance Year Developmental, 2010 Principal Investigators, et al. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010. *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, 63:1, 2014.

[55] Disabilities Monitoring Network Surveillance Year Developmental, 2010 Principal Investigators, et al. Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010. *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, 63:1, 2014.

[56] Joshua J Diehl, Lauren M Schmitt, Michael Villano, and Charles R Crowell. The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Research in autism spectrum disorders*, 6(1):249–262, 2012.

[57] Sidney DMello, Rosalind Picard, and Arthur Graesser. Towards an affect-sensitive autotutor. *IEEE Intelligent Systems*, 22(4):53–61, 2007.

[58] Dr Guillaume Benjamin Duchenne. *Mécanisme de la physionomie humaine...* typ. F. Malteste, 1862.

[59] GB Duchenne de Bologne. Mechanisme de la physionomie humaine. *Jules Renouard Libraire. Paris*, 1862.

[60] Starkey Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

[61] Audrey Duquette, François Michaud, and Henri Mercier. Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Autonomous Robots*, 24(2):147–157, 2008.

[62] P Ekman and W Friesen. Rationale and reliability for emfacs coders. *Unpublished manuscript*, 1982.

[63] Paul Ekman and Wallace V Friesen. Facial action coding system. 1977.

[64] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture*, pages 57–106, 1981.

[65] Paul Ekman, Wallace V Friesen, and Joseph C Hager. *Facial action coding system*. A Human Face Salt Lake City, 2002.

[66] Paul Ekman, Wallace V Friesen, and Joseph C Hager. *Facial action coding system*. A Human Face Salt Lake City, 2002.

[67] Paul Ekman, Wallace V Friesen, and Silvan S Tomkins. Facial affect scoring technique: A first validity study. *Semiotica*, 3(1):37–58, 1971.

[68] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[69] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[70] Beat Fasel and Juergen Luettin. Recognition of asymmetric facial action unit activities and intensities. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 1100–1103. IEEE, 2000.

[71] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[72] David Feil-Seifer and Maja Mataric. Robot-assisted therapy for children with autism spectrum disorders. In *Proceedings of the 7th international conference on Interaction design and children*, pages 49–52. ACM, 2008.

[73] Anne Fernald and Claudia Mazzie. Prosody and focus in speech to infants and adults. *Developmental psychology*, 27(2):209, 1991.

[74] M Frank, J Movellan, M Bartlett, and G Littleworth. Ru-facs-1 database. *Machine Perception Laboratory, UC San Diego*, 1, 2012.

[75] Dario Galati, Barbara Sini, Susanne Schmidt, and Carla Tinti. Spontaneous facial expressions in congenitally blind and sighted children aged 8-11. *Journal of Visual Impairment & Blindness*, 97(7), 2003.

[76] Haisong Gu and Qiang Ji. Information extraction from image sequences of real-world facial expressions. *Machine Vision and Applications*, 16(2):105–115, 2005.

[77] Jihun Hamm, Christian G Kohler, Ruben C Gur, and Ragini Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2):237–256, 2011.

[78] Zakia Hammal, Jeffrey F Cohn, Daniel S Messinger, Whitney I Mattson, and Mohammad H Mahoor. Head movement dynamics during normal and perturbed parent-

infant interaction. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 276–282. IEEE, 2013.

[79] David Hanson, Stuart Baurmann, Thomas Riccio, Richard Margolin, Tim Dockins, Matthew Tavares, and Kevin Carpenter. Zeno: A cognitive character. In *AI Magazine, and special Proc. of AAAI National Conference, Chicago*, 2009.

[80] Catherine Herba and Mary Phillips. Annotation: Development of facial expression recognition from childhood to adolescence: Behavioural and neurological perspectives. *Journal of Child Psychology and Psychiatry*, 45(7):1185–1198, 2004.

[81] Carl-Herman Hjortsjö. *Man's face and mimic language*. Studentlitteratur, 1969.

[82] Jean-Michel Hoc. From human–machine interaction to human–machine cooperation. *Ergonomics*, 43(7):833–843, 2000.

[83] Amanda Holmes, Patrik Vuilleumier, and Martin Eimer. The processing of emotional facial expression is gated by spatial attention: evidence from event-related brain potentials. *Cognitive Brain Research*, 16(2):174–184, 2003.

[84] VANESSA Hugot. Eye gaze analysis in human-human interactions. *Unpublished masters thesis, CSC, KTH, Sweden*, 2007.

[85] Carroll Ellis Izard. *The Maximally Discriminitive Facial Movements Coding System, MAX*. University of Delaware, 1979.

[86] Carroll Ellis Izard, Linda M Dougherty, and Elizabeth Ann Hembree. *A system for identifying affect expressions by holistic judgments (AFFEX)*. Instructional Resources Center, University of Delaware, 1983.

[87] Biing Hwang Juang and Laurence R Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[88] Susanne Kaiser and Thomas Wehrle. Automated coding of facial behavior in human-computer interactions with facs. *Journal of Nonverbal Behavior*, 16(2):67–84, 1992.

[89] Takeo Kanade, Jeffrey F Cohn, and Yingli Tian. Comprehensive database for facial expression analysis. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 46–53. IEEE, 2000.

[90] Dacher Keltner, Paul Ekman, Gian C Gonzaga, and Jennifer Beer. Facial expression of emotion. 2003.

[91] Elizabeth S Kim, Lauren D Berkovits, Emily P Bernier, Dan Leyzberg, Frederick Shic, Rhea Paul, and Brian Scassellati. Social robots as embedded reinforcers of social behavior in children with autism. *Journal of autism and developmental disorders*, 43(5):1038–1049, 2013.

[92] Ami Klin. Autism and asperger syndrome: an overview. *Revista brasileira de psiquiatria*, 28:s3–s11, 2006.

[93] Irene Kotsia, Ioan Buciu, and Ioannis Pitas. An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, 26(7):1052–1067, 2008.

[94] Hideki Kozima, Cocoro Nakagawa, and Yuriko Yasuda. Interactive robots for communication-care: A case-study in autism therapy. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pages 341–346. IEEE, 2005.

[95] Stephen RH Langton. The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A*, 53(3):825–845, 2000.

[96] Yongqiang Li, S Mohammad Mavadati, Mohammad H Mahoor, and Qiang Ji. A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013.

[97] James Jenn-Jier Lien, Takeo Kanade, Jeffrey F Cohn, and Ching-Chung Li. Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, 31(3):131–146, 2000.

[98] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[99] Gwen C Littlewort, Marian Stewart Bartlett, and Kang Lee. Faces of pain: automated measurement of spontaneousallfacial expressions of genuine and posed pain. In *Proceedings of the 9th international conference on Multimodal interfaces*, pages 15–21. ACM, 2007.

[100] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981.

[101] P Lucey, J Cohn, J Howlett, S Lucey, and S Sridharan. Recognizing emotion with head pose variation: Identifying pain segments in video. *IEEE Trans. Syst. Man Cybern., Part B, Cybern*, 41(3):664–674, 2011.

[102] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain 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.

[103] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain 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.

[104] Simon Lucey, Ahmed Bilal Ashraf, and Jeffrey Cohn. Investigating spontaneous facial action recognition through aam representations of the face. *Face recognition*, pages 275–286, 2007.

[105] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro 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.

[106] Mohammad H Mahoor, Steven Cadavid, Daniel S Messinger, and Jeffrey 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.

[107] Mohammad H Mahoor, Mu Zhou, Kevin L Veon, Seyed Mohammad Mavadati, and Jeffrey F Cohn. Facial action unit recognition with sparse representation. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 336–342. IEEE, 2011.

[108] Mohammad Hossein Mahoor and Mohamed Abdel-Mottaleb. A multimodal approach for face modeling and recognition. *Information Forensics and Security, IEEE Transactions on*, 3(3):431–440, 2008.

[109] Carol Z Malatesta, Clayton Culver, Johanna R Tesman, and Beth Shepard. The development of emotion expression during the first two years of life. *Monographs of the Society for Research in Child Development*, 1989.

[110] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, and Philip Trinh. Automatic detection of non-posed facial action units. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 1817–1820. IEEE, 2012.

[111] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, pages 151–160, 2013.

[112] Erin B McClure. A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents. *Psychological bulletin*, 126(3):424, 2000.

[113] Pauline Mendola, Sherry G Selevan, Suzanne Gutter, and Deborah Rice. Environmental factors associated with a spectrum of neurodevelopmental deficits. *Mental retardation and developmental disabilities research reviews*, 8(3):188–197, 2002.

[114] Daniel S Messinger, Mohammad H Mahoor, Sy-Miin Chow, and Jeffrey F Cohn. Automated measurement of facial expression in infant–mother interaction: A pilot study. *Infancy*, 14(3):285–305, 2009.

[115] Daniel S Messinger, Whitney I Mattson, Mohammad H Mahoor, and Jeffrey F Cohn. The eyes have it: Making positive expressions more positive and negative expressions more negative. *Emotion*, 12(3):430, 2012.

[116] Philipp Michel and Rana El Kaliouby. Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on Multimodal interfaces*, pages 258–264. ACM, 2003.

[117] Vanessa Moore and Sally Goodson. How well does early diagnosis of autism stand the test of time? follow-up study of children assessed for autism at age 2 and development of an early diagnostic service. *Autism*, 7(1):47–63, 2003.

[118] John Morton and Mark H Johnson. Conspec and conlern: a two-process theory of infant face recognition. *Psychological review*, 98(2):164, 1991.

[119] SL Odom. Technology-aided instruction and intervention (taii) fact sheet. *Chapel Hill (NC): The University of North Carolina*, 2013.

[120] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[121] Harriet Oster, Douglas Hegley, and Linda Nagel. Adult judgments and fine-grained analysis of infant facial expressions: Testing the validity of a priori coding formulas. *Developmental Psychology*, 28(6):1115, 1992.

[122] Maja Pantic and Ioannis Patras. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, volume 4, pages 3358–3363. IEEE, 2005.

[123] Maja Pantic and Ioannis Patras. Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 36(2):433–449, 2006.

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

[125] Maja Pantic and Leon J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424–1445, 2000.

[126] Maja Pantic and Leon JM Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[127] Maja Pantic and Leon JM Rothkrantz. Facial action recognition for facial expression analysis from static face images. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 34(3):1449–1461, 2004.

[128] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 5–pp. IEEE, 2005.

[129] DI Perrett, JK Hietanen, MW Oram, PJ Benson, and ET Rolls. Organization and functions of cells responsive to faces in the temporal cortex [and discussion]. *Philosophical transactions of the royal society of London. Series B: Biological sciences*, 335(1273):23–30, 1992.

[130] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[131] Kenneth M Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51(3):297–306, 1992.

[132] Nadja Reissland, Brian Francis, James Mason, and Karen Lincoln. Do facial expressions develop before birth? *PLoS One*, 6(8):e24081, 2011.

[133] Daniel J Ricks and Mark B Colton. Trends and considerations in robot-assisted autism therapy. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 4354–4359. IEEE, 2010.

[134] Ben Robins, Kerstin Dautenhahn, and Janek Dubowski. Does appearance matter in the interaction of children with autism with a humanoid robot? *Interaction Studies*, 7(3):509–542, 2006.

[135] Sally J Rogers. Brief report: Early intervention in autism. *Journal of autism and developmental disorders*, 26(2):243–246, 1996.

[136] Yvonne Rogers, Helen Sharp, and Jenny Preece. *Interaction design: beyond human-computer interaction*. John Wiley & Sons, 2011.

[137] Georgia Sandbach, Stefanos Zafeiriou, and Maja Pantic. Markov random field structures for facial action unit intensity estimation.

[138] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *Biometrics and Identity Management*, pages 47–56. Springer, 2008.

[139] Michael A Sayette, Jeffrey F Cohn, Joan M Wertz, Michael A Perrott, and Dominic J Parrott. A psychometric evaluation of the facial action coding system for assessing spontaneous expression. *Journal of Nonverbal Behavior*, 25(3):167–185, 2001.

[140] Brian Scassellati, Henny Admoni, and Maja Mataric. Robots for use in autism research. *Annual Review of Biomedical Engineering*, 14:275–294, 2012.

[141] Karen L Schmidt, Zara Ambadar, Jeffrey F Cohn, and L Ian Reed. Movement differences between deliberate and spontaneous facial expressions: Zygomaticus major action in smiling. *Journal of Nonverbal Behavior*, 30(1):37–52, 2006.

[142] Miriam Schmidt, Martin Schels, and Friedhelm Schwenker. A hidden markov model based approach for facial expression recognition in image sequences. In *Artificial Neural Networks in Pattern Recognition*, pages 149–160. Springer, 2010.

[143] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[144] Frederick Shic, Katarzyna Chawarska, Jessica Bradshaw, and Brian Scassellati. Autism, eye-tracking, entropy. In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, pages 73–78. IEEE, 2008.

[145] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[146] Chang Shu, Xiaoqing Ding, and Chi Fang. Histogram of the oriented gradient for face recognition. *Tsinghua Science & Technology*, 16(2):216–224, 2011.

[147] Mohammad H. Mahoor S.Mohammad Mavadati. Temporal facial expression modeling for automated action unit intensity measurement. *International Conference on Pattern Recognition*, 2014.

[148] Ian Sneddon, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. The belfast induced natural emotion database. *Affective Computing, IEEE Transactions on*, 3(1):32–41, 2012.

[149] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *AI 2006: Advances in Artificial Intelligence*, pages 1015–1021. Springer, 2006.

[150] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115, 2001.

[151] Ying-Li Tian, Takeo Kanade, and Jeffrey F Cohn. Facial expression analysis. In *Handbook of face recognition*, pages 247–275. Springer, 2005.

[152] Yan Tong, Jixu Chen, and Qiang Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(2):258–273, 2010.

[153] Colwyn Trevarthen. Intrinsic motives for companionship in understanding: Their origin, development, and significance for infant mental health. *Infant Mental health journal*, 22(1-2):95–131, 2001.

[154] Michel F Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM, 2006.

[155] Michel François Valstar. *Timing is everything: A spatio-temporal approach to the analysis of facial actions*. PhD thesis, Imperial College London, 2008.

[156] Vladimir N Vapnik. An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5):988–999, 1999.

[157] S Vicari, J Snitzer Reilly, P Pasqualetti, A Vizzotto, and C Caltagirone. Recognition of facial expressions of emotions in school-age children: the intersection of perceptual and semantic categories. *Acta Paediatrica*, 89(7):836–845, 2000.

[158] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[159] Javier Virués-Ortega. Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical psychology review*, 30(4):387–399, 2010.

[160] Jun Wang and Lijun Yin. Static topographic modeling for facial expression recognition and analysis. *Computer Vision and Image Understanding*, 108(1):19–34, 2007.

[161] Shangfei Wang, Zhilei Liu, Siliang Lv, Yanpeng Lv, Guobing Wu, Peng Peng, Fei Chen, and Xufa Wang. A natural visible and infrared facial expression database for expression recognition and emotion inference. *Multimedia, IEEE Transactions on*, 12(7):682–691, 2010.

[162] Zachary E Warren, Zhi Zheng, Amy R Swanson, Esubalew Bekele, Lian Zhang, Julie A Crittendon, Amy F Weitlauf, and Nilanjan Sarkar. Can robotic interaction improve joint attention skills? *Journal of autism and developmental disorders*, pages 1–9, 2013.

[163] Sylvia Weir and Ricky Emanuel. *Using LOGO to catalyse communication in an autistic child*. Department of Artificial Intelligence, University of Edinburgh, 1976.

[164] Jacob Whitehill, Marian Bartlett, and Javier Movellan. Automatic facial expression recognition for intelligent tutoring systems. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–6. IEEE, 2008.

[165] Jacob Whitehill, Gwen Littlewort, Ian Fasel, Marian Bartlett, and Javier Movellan. Toward practical smile detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2106–2111, 2009.

[166] Amanda C de C Williams et al. Facial expression of pain: an evolutionary account. *Behavioral and brain sciences*, 25(4):439–455, 2002.

[167] C Wong, SL Odom, K Hume, AW Cox, A Fetting, S Kucharczyk, and TR Schultz. Evidence-based practices for children, youth, and young adults with autism spectrum disorder. *The University of North Carolina, Frank Porter Graham Child Development Institute, Autism Evidence-Based Practice Review Group, Chapel Hill, NC*, 2013.

[168] Connie Wong, Samuel L Odom, Kara Hume, Ann W Cox, Angel Fettig, Suzanne Kucharczyk, Matthew E Brock, Joshua B Plavnick, Veronica P Fleury, and Tia R Schultz. Autism spectrum disorder. 2014.

[169] S.M. Mavadati X. Zhang, M.H. Mahoor and J.F. Cohn. A lp-norm mtmkl framework for simultaneous detection of multiple facial action units. *IEEE Winter conference on Applications of Computer Vision (WACV14),*, 2014.

[170] Xuehan Xiong and Fernando 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.

[171] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[172] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*, pages 211–216. IEEE, 2006.

[173] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.

[174] Yongmian Zhang, Qiang Ji, Zhiwei Zhu, and Beifang Yi. Dynamic facial expression analysis and synthesis with mpeg-4 facial animation parameters. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(10):1383–1396, 2008.

[175] Yunfeng Zhu, Fernando De la Torre, Jeffrey F Cohn, and Yu-Jin Zhang. Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *Affective Computing, IEEE Transactions on*, 2(2):79–91, 2011.

Appendix

A.1 Hidden Markov Model (HMM)

Hidden Markov Model is a powerful stochastic state model which is utilized to represent and classify time-sequential data. The discrete HMM has a with set of unobservable (hidden) states that at each time t, emits one output from a group of observable symbols. A (homogeneous) HMM can be represented by N states, $Q = q_1, q_2, , q_N$, and state transition probability $A = [a_{ij}] \in R^{N \times N}$ can be formulated as:

$$\begin{aligned} a_{ij} = Pr(s_{t+1}{=}q_j | s_t{=}q_i), \quad 1 \leq i, j \leq N, \\ s.t.\ \ a_{ij} \geq 0,\ \ \sum_{j=1}^{N} a_{ij} = 1 \end{aligned} \tag{A.1.1}$$

Moreover, the HMM has a set of M observable symbols $V = [v_1, v_2, , v_M]$ that is emitted from one of the hidden states. Emission probability (i.e. output probability) is probability of producing an output symbol v_k being in the state q_j.

$$b_j(k){=}Pr(v_k | s_t{=}q_j), \quad 1{\leq}j{\leq}N,\ 1{\leq}k{\leq}M,\ 1{\leq}t{\leq}T \tag{A.1.2}$$

A HMM is characterized by state transition probability matrix (A), emission probability (B), and initial state probability (π). Hidden Markov model $\lambda = (A, B, \pi)$ enable us to model a time sequence of data with length T $(t = 1, 2, \ldots, T)$ that can be applied for three types of applications as follows.

- **Decoding:** Given parameters of a HMM, $\lambda = (A, B, \pi)$, and a sequence of an observed output $O = \{O_1, O_2, \ldots, O_T\}$, evaluate the most probable sequence of states $S = s_t, t = 1, 2, \ldots, T$ for generating the sequence of outputs.

- **Evaluation:** Given HMM parameters, specify the probability of an observed output sequence $O = \{O_1, O_2, \ldots, O_T\}$.
- **Learning:** Given the structure of HMM model (i.e. states and observations) and a set of output sequences $O = \{O_1, O_2, \ldots, O_T\}$, estimate the HMM parameters $\lambda = (A, B, \pi)$.

In order to exploit HMM for classification applications, both evaluation and learning phases need to be considered [171]. This has been explained in Section A.1.1.

A.1.1 Classifying time-sequential data

To classify an observed sequence of symbols $\{O = O_1, O_2, \ldots, O_T\}$ to a category, we create one HMM for each category. In the *Evaluation Phase*, we aim to find one HMM which best matches the observation sequence O. Mathematically speaking, for the sequence $\{O\}$ we calculate $Pr(\lambda_i|O)$ for all C HMMs and select λ_{C^*}, where

$$C^* = \arg\max(Pr(\lambda_i|O)) \tag{A.1.3}$$

To compute this optimization efficiently, logarithm of the probabilities will be utilized which results in the log-likelihood estimation. The *Forward Induction Algorithm* is one of the techniques that can efficiently solve this problem . In forwards algorithm a *forward variable* $\alpha_t(i)$ will be defined as:

$$\alpha_t(i) = Pr(O_1, O_2, \ldots, O_N, q_t = i \mid \lambda) \tag{A.1.4}$$

where $\alpha_t(i)$ calculates the probability of the partial sequence of observation at time t being in state $q_t = i$. We can exploit this definition to recursively calculate $\alpha_t(i)$ as follows.

$$\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} \right] b_j(O_t) \tag{A.1.5}$$

Equation A.1.5 can significantly reduce the computational cost of optimization (i.e. linear in the sequence length T and quadratic in number of states N). Consequently the results for $Pr(O|\lambda)$ can be calculated by

$$Pr(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \tag{A.1.6}$$

In order to find the best HMM parameters, $\lambda = (A, B, \pi)$, for classifying C classes of data *learning algorithm* need to be introduced. In the learning phase, C HMMs will be trained and HMM parameters for one class $\{C_i; i = 1, 2, \ldots, C\}$ will be trained to achieve the highest likelihood value for the corresponding observed sequence of symbols. In other words, for any given HMM parameter (A, B, π) we maximize the observation sequence probability $Pr(O|\lambda)$. In order to find right model parameter value the Buam-Welch algorithm can be used which maximize the objective function in a two-step procedure [171]. In the first step the objective function $Pr(O|\lambda)$ will be transformed to $Q(\lambda^{'}, \lambda)$ as defined bellow:

$$Q(\lambda^{'}, \lambda) = \sum_{q} Pr(O, q|\lambda^{'}) \; logPr(O, q|\lambda)) \tag{A.1.7}$$

where $Q(\lambda^{'}, \lambda)$ can measure the divergence of initial guess $\lambda^{'}$ w.r.t the updated model parameter λ. As $Q(\lambda^{'}, \lambda) \geq Q(\lambda^{'}, \lambda^{'})$ it implies that $Pr(O|\lambda^{'}) \leq Pr(O|\lambda)$ and we can maximize $Q(\lambda^{'}, \lambda)$ over λ. In the second step of algorithm $Q(\lambda^{'}, \lambda)$ can be updated using Expectation Maximization(EM) technique. This would lead to finding the right model parameter for a HMM. More details about evaluation and learning phases can be found in [171].

A.1.2 HMM and video sequence recognition

In order to employ HMM for classifying a video or sequence of images, $I = \{I_1, I_2, \ldots, I_T\}$, every image needs to be represented by a symbol (O_i). Then using the sequence of symbols $O = \{O_1, \ldots, O_T\}$ for both learning and recognition phases, will enable us to categorize the video segment with length T into one of C classes.

For an image I_i $(i = 1, \ldots, T)$ a feature vector $f_i \in R^n$ represents the image in *n*-dimensional feature space . Afterwards the extracted feature needs to be assigned to one of the outputs from the symbol set V. To associate the symbol set V with feature space f we utilized vector quantization technique to divide *n*-dimensional space into M clusters. K-means clustering is a powerful technique to partition n-dimensional data into M clusters by finding local minimum to the distance function:

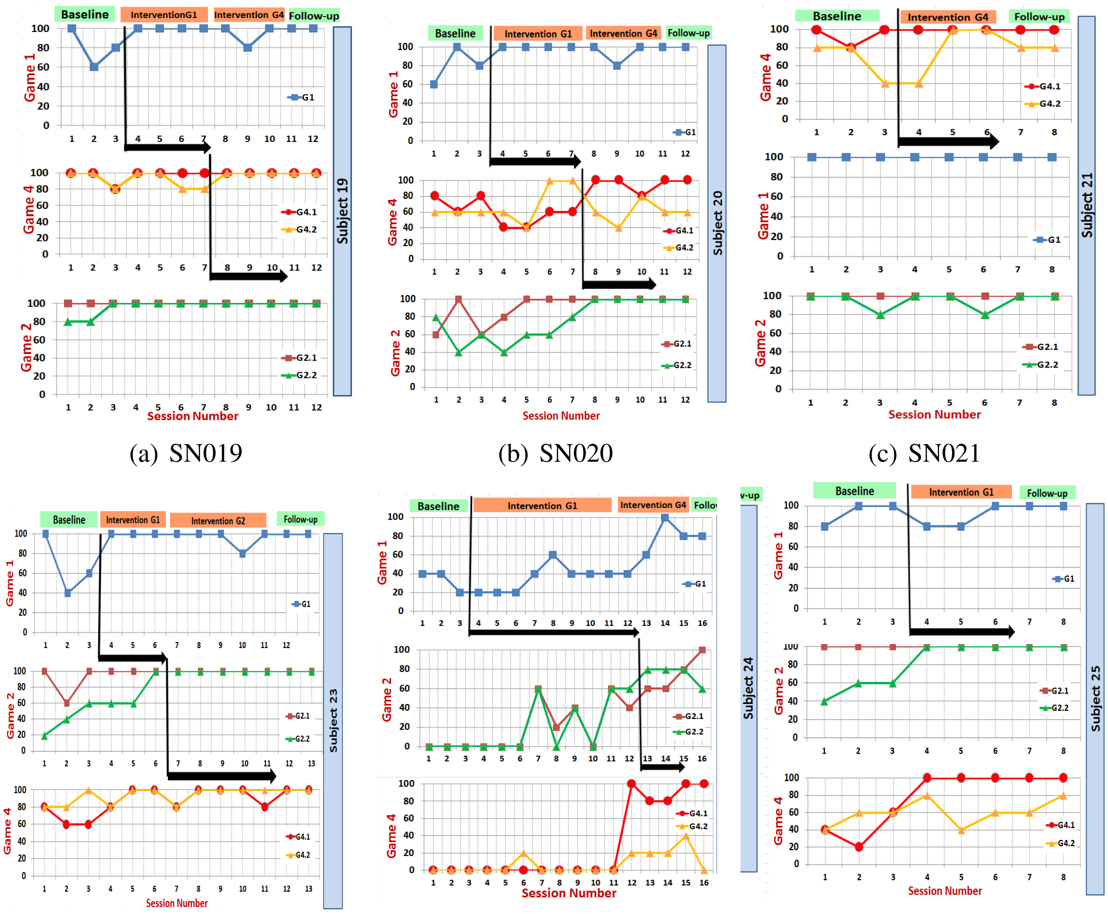
$$J(g_1, g_2, \ldots, g_M) = \sum_{i=1}^{C} \sum_{j \in C_j} ||f_j - g_i||_2^2 \tag{A.1.8}$$

where f_j represent the data that belongs to the C_J category and g_i is the centroid of i^{th} partition. The centroid of i^{th} cluster or g_i will be assigned to symbol v_i. This procedure will help us to design a codebook that comprises of M number of symbols for HMM. Therefore each feature vector f_j will be assigned into the nearest codeword g_h where $h = \arg\min dist(g_i, f_j)$ for $(i = 1, \ldots, M)$.

As described in section A.1.1, to learn HMM parameters of a category, a set of data that belongs to this category (i^{th} class), will be employed to optimize the HMM parameters (i.e. $\lambda_i(A_i, B_i, \pi_i)$). Furthermore, to recognizing the category of an observed sequence symbols ,O, the $Pr(\lambda_i|O); i = 1, \ldots, C$ will be calculated and the observed sequence will be assigned to a class which has the highest likelihood probability.

A.2 Autism and Robot-based Intervention Results

For our robot-based therapeutic setting seven children with ASD were recruited and they participated in multi-session interactions with NAO. Considering the baseline data, all subjects demonstrated high level of accuracy for two games (i.e. G3 and G5). Therefore Figure A.1 illustrated the behavioral and social responses of six children (whom needed interventions) for the remaining one or tow games out of the three games that the child may required intervention (i.e. G1, G2, and G4). In Figure A.2, the vertical line demonstrates the starting session of each intervention and horizontal arrow specify the duration of the intervention sessions for each game (for other sessions no feedback were provided by NAO).



(a) SN019 (b) SN020 (c) SN021

(d) SN023 (e) SN024 (f) SN025

Figure A.1: Baseline, intervention and follow-up sessions

A.3 My Publications

Journal Papers:

1. **S.M. Mavadati**, M.H. Mahoor, K. Bartlett, P. Trinh, J. Cohn, *“DISFA: A Spontaneous Facial Action Intensity Database”*, Transaction on Affective Computing IEEE, vol.4, no.2, pp.151,160, April-June 2013
2. **S.M. Mavadati**, H. Feng, A. Gutierrez, M.H. Mahoor, *“How Children with Autism Regulate Their Eye Gaze in Interaction with a Robot?”*, Journal of Autism and Developmental Disorders (JADD, 2014)– Under Review.
3. **S.M. Mavadati**, M. Salvador, H. Feng, S. Silver, A. Gutierrez, M.H. Mahoor, *“Pilot Study: Robot-based Behavioral Intervention for Individuals with Autism Spectrum Disorder ”*, Journal of Autism and Developmental Disorders (JADD, 2015)– Under Review.
4. J.M. Girard, J.F. Cohn, M.M Mahoor, **S.M. Mavadati**, Z. Harnmal, D. P. Rosenwald, *Social Risk and Depression: Evidence from Nonverbal Analysis”*, Image and Vision Computing Journal (IVCJ2014).
5. Y. Li, **S.M. Mavadati**, M. H. Mahoor, Y. Zhao, Q. Ji, *“Measuring the Intensity of Spontaneous Facial Action Units with Dynamic Bayesian Network”*, Pattern Recognition - Journal (Elsevier)–Under Review.
6. X. Zhang, M. H. Mahoor, **S.M. Mavadati**, *”A lp-norm MKL Multiclass-SVM Framework for Facial Expression Recognition”*, Machine Vision and Applications Journal (MVA)– Under Review.

Refereed Conference Papers:

7. **S.M. Mavadati**, H. Feng, P. B. Sanger, S. Silver, A. Gutierrez, M. H. Mahoor, *”Using Robots As Therapeutic Agents to Teach Children with Autism Recognize Facial Expression”*, International Meeting for Autism Research (IMFAR 2015).

8. **S.M. Mavadati**, H. Feng, A. Gutierrez, M. H. Mahoor, *"Comparing the Gaze Responses of Children with Autism and Typically Developed Individuals in Human-Robot Interaction"*, International Conference on Humanoid Robots (Humanoids 2014).
9. **S.M. Mavadati**, M. H. Mahoor , *"Temporal Facial Expression Modeling for Automated Action Unit Intensity Measurement"*, International Conference on Pattern Recognition (ICPR 2014).
10. **S.M. Mavadati**, H. Feng, S. Silver, A. Gutierrez, M. H. Mahoor, *"Children-Robot Interaction: Eye Gaze Analysis of Children with Autism during Social Interaction"*, International Meeting for Autism Research (IMFAR 2014).
11. **S.M. Mavadati**, M. H. Mahoor, X. Zhang, *"Manifold Alignment Using Curvature Information"*, International Conference on Image and Vision Computing New Zealand (IVCNZ 2013).
12. **S.M. Mavadati**, M.H. Mahoor, K. Bartlett, P. Trinh, *"Automatic Detection of Non-posed Facial Action Units"*, International Conference in Image Processing (ICIP 2012).
13. Y. Li,**S.M. Mavadati**, M.H. Mahoor, Q. Ji, *"A Unified Probabilistic Framework For Measuring The Intensity of Spontaneous Facial Action Units"*, IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013), Shanghai, China. April 2013.
14. J.M. Girard, J.F. Cohn, M.H. Mahoor, **S.M. Mavadati**, D. Rosenwald, and F.D.L. Torre. *"Manual and automatic analysis of facial affective reactivity in major depressive disorder"*. IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013), Shanghai, China. April 2013.
15. **S.M. Mavadati**, M.H. Mahoor, *"A New Approach for Curvature Estimation of Sampled Data"*, International Conference on Development and Learning (ICDL 2012), San Diego, CA, November 2012.