The University of Nottingham
Faculty of Science
School of Computer Science

# Dynamic Deep Learning for Automatic Facial Expression Recognition and its Application in Diagnosis of ADHD & ASD

Shashank Jaiswal

Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy in Computer Science of The University of Nottingham. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

January 2018

## Abstract

Neurodevelopmental conditions like Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD) impact a significant number of children and adults worldwide. Currently, the means of diagnosing of such conditions is carried out by experts, who employ standard questionnaires and look for certain behavioural markers through manual observation. Such methods are not only subjective, difficult to repeat, and costly but also extremely time consuming. However, with the recent surge of research into automatic facial behaviour analysis and it's varied applications, it could prove to be a potential way of tackling these diagnostic difficulties. Automatic facial expression recognition is one of the core components of this field but it has always been challenging to do it accurately in an unconstrained environment. This thesis presents a dynamic deep learning framework for robust automatic facial expression recognition. It also proposes an approach to apply this method for facial behaviour analysis which can help in the diagnosis of conditions like ADHD and ASD.

The proposed facial expression algorithm uses a deep Convolutional Neural Networks (CNN) to learn models of facial Action Units (AU). It attempts to model three main distinguishing features of AUs: shape, appearance and short term dynamics, jointly in a CNN. The appearance is modelled through local image regions relevant to each AU, shape is encoded using binary masks computed from automatically detected facial landmarks and dynamics is encoded by using a short sequence of image as input to CNN. In addition, the method also employs Bi-directional Long Short Memory (BLSTM) recurrent neural networks for modelling long term dynamics. The proposed approach is evaluated on a number of databases showing state-of-the-art performance for both AU detection and intensity estimation tasks. The AU intensities estimated using this approach along with other 3D face tracking data, are used for encoding facial behaviour. The encoded facial behaviour is applied for learning models which can help in detection of ADHD and ASD. This approach was evaluated on the KOMAA database which was specially collected for this purpose. Experimental results show that facial behaviour encoded in this way provide a high discriminative power for classification of people with these conditions. It is shown that the proposed system is a potentially useful, objective and time saving contribution to the clinical diagnosis of ADHD and ASD.

**Acknowledgements**

I would like to express my thanks to my supervisor Dr. Michel Valstar for giving me the opportunity to work in this exciting area and for his guidance and support during the entire duration of my PhD. I would also like to thank my second supervisor Prof. David Daley for his guidance on the mental health aspect of my research. This thesis would not have been possible without the endless love and support of my family who have always encouraged me to follow my passion. I am grateful to them for their unconditional love and support whenever I needed. A big and special thanks goes to my partner Catrin who supported me tremendously in whatever I did. It would be difficult to imagine this thesis without her love and support. Finally, I would also like to thank all my friends and colleagues at the School of Computer Science, especially, Andry, Timur, Amy, Tuan, Enrique, Xiaomeng, Alex, Matthew, Kelvin, Hyosun and Leigh. It was a pleasure to work in such a friendly and creative atmosphere.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Human behaviour comprises a set of visual communicative signals through which a person expresses himself/herself. It acts as a window to an individual's thoughts and feelings and provides an insight into a person's psyche. It is for this reason that the analysis of human behaviour can be extremely useful in the fields of psychology, mental health, human-computer interaction, etc. For example, mental disorders such as depression and Attention Deficit Hyperactivity Disorder (ADHD) directly influence an individual's behavior in his/her day to day life. Analysis of human behaviour is often used in the diagnosis and treatment of such disorders. Similarly, many elements of human behaviour including gestures and facial expressions play an important role in the way humans communicate with each other. Understanding the role of human behaviour in the way humans communicate with each other can help in designing new machine interfaces which make the human-machine interaction more naturalistic and easier for humans.

A major part of the non-verbal behavioural cues is in the form of facial gestures and expressions. They are also the one of the most widely studied parts of human behaviour. The internal emotional state of a person directly influences (sometimes unconsciously) the movement of facial muscles resulting in different facial expressions and head gestures. For example, an emotional state of 'surprise' results in the raising of eyebrows and opening of the mouth. Similarly, the emotional state of 'happiness' is often accompanied by the pulling of lip corners (smiles). These facial muscle movements not only act a window to an individual's internal emotional state and intentions but they also form a significant part of human to human interaction, using a non-verbal mode of communication. Therefore, study of such non-verbal facial cues is of paramount importance for

advancement in the field of human psychology and social-signal processing.

Interest in studying facial expressions started in the nineteenth century when Darwin (Darwin et al., 1998) suggested that certain basic emotions are expressed in a remarkably similar way across cultures and species. He argued that facial expressions are inherited and are a result of the process of evolution. In 1978, Ekman and Fiesen(Ekman et al., 2002) developed the Facial Action Coding System (FACS). It provided a systematic and objective way to study facial expressions by representing them as a combination of individual or a group of facial muscle actions known as Action Units (AU).

Human beings can easily recognize each other's facial gestures and expressions. However, in order to develop automated systems whose interfaces have the ability to understand and respond to non-verbal cues such as facial expressions, it is necessary to develop automatic facial expression and gesture recognition algorithms. The development of such algorithms becomes even more important if we want to collect a large amount of data for a deeper understanding of such facial cues and the role they can play in the field of psychology and other related areas. Due to such potential applications, automatic facial expression recognition algorithms have attracted considerable attention in the past few years (Sandbach et al., 2012; Sariyanidi et al., 2015; Corneanu et al., 2016; Zafeiriou et al., 2016). The recent advancement in the field of machine learning and computer vision has made it possible to automatically recognize facial expressions more accurately than ever before. However, the problem of automatic facial expression recognition in unconstrained environments is still a challenging problem, and is far from solved. It is primarily due to the high degree of variation in the visual appearance of human faces (caused by person specific attributes, multiple poses, etc.), and non-availability of large amount of labelled training data.

Recently, deep learning algorithms have shown significant performance improvements for object detection tasks (Krizhevsky et al., 2012a; Girshick et al., 2014). The recent success of deep learning algorithms has been attributed to three main factors: (a) Efficient methods for training deep networks, (b) Availability of high performance computational hardware e.g. GPUs (c) Availability of large amounts of labeled training data. Although deep learning algorithms have been shown to produce state of the art performance on object recognition tasks, there has been considerably less work on using deep learning techniques in facial expression recognition and in particular facial AU recognition. With the increasing availability of large databases for AU recognition (Mava-

dati et al., 2013; Valstar et al., 2015a), it would be interesting to see if deep learning algorithms can give a similar leap in performance in the field of facial expression/AU recognition.

Traditional AU recognition algorithms have used hand crafted appearance features (e.g. HoG, LBP) and/or shape features computed from the locations of facial landmarks. Since these hand crafted features are not tuned to the specific task at hand, they limit the performance of the classifier learnt on these features. Deep learning techniques on the other hand incorporate a multistage technique in which the features are learnt directly from the pixel values in combination with the classifier. Therefore, in addition to providing an algorithm which can be trained directly from pixels to labels, the features learnt in the intermediate stages are designed specifically for the target task.

In this thesis, the aim is to explore the application of neural network architectures, with specific focus on deep learning architectures, for automatic facial expression/AU recognition. The performance of the existing ANN algorithms is assessed and new ANN algorithms have been explored for improved facial expression/AU recognition. Specifically we focus more towards Convolutional Neural Networks (CNN) architectures as these are currently the state of the art for object detection tasks. We also focus on recurrent neural network architectures such as Long Short-Term Memory (LSTM) which are specialized for modeling temporal information.

Real world application of automatic facial expression recognition is another aspect that we explore in this thesis. The steady progress in automatic expressive behaviour analysis in the last 5 years: detection and tracking of faces (Zhu and Ramanan, 2012b; Mathias et al., 2014; Xiong and De la Torre Frade, 2013), recognition of facial muscle actions (Valstar and Pantic, 2012b; Chu et al., 2013a; Valstar et al., 2015b), and accurate head pose estimation (Zhu and Ramanan, 2012b; Yan et al., 2016), has renewed the interest of researchers in employing such behaviour analysis in the medical domain, targeting so-called *behaviomedical* conditions that alter one's expressive behaviour (Valstar, 2014). In particular, this thesis explores the feasibility of using automatic facial expression and gesture analysis for the diagnosis of Attention Deficit Hyperactivity disorder (ADHD) and Autism Spectrum Disorder (ASD). Both ADHD and ASD are neurodevelopmental conditions which impact a significant number of people. It has been estimated that at least 2.5% of the population (Simon et al., 2009) is affected by ADHD, which is characterized by symptoms such as inattention, hyperactivity, impulsivity, etc. (Barkley, 1997;

Spencer et al., 2007). It usually begins in early childhood and quite often the symptoms persists into adulthood (Geissler and Lesch, 2011). It is widely believed that both genetic (Stevenson et al., 2005) and environmental influences (Larsson et al., 2004) contribute to the underlying cause of this disorder. Presently, the diagnosis of ADHD is made following the criteria of the DSM-5 (APA, 2013), which involve mechanisms to validate hyperactivity, attention deficit and impulsivity. The diagnosis is made by experts using a combination of developmental history, collateral information, psychometrics and behavioural observation and impairment. This is often difficult and time consuming.

ADHD is also known to show co-morbidity with ASD (Autism Spectrum Disorder). ASD is a neuro-developmental condition which is characterised by impairments in social interaction and communication and restricted, repetitive or stereotyped behaviours and interests (Faras et al., 2010). It has been found that a significant number of people with ASD also show symptoms of ADHD (Rao and Landa, 2013). Treatment methods also vary for all 3 groups of people i.e. only ASD, only ADHD, and ADHD+ASD. Hence, accurate diagnosis can have important implications for treatment. However, currently the manual diagnosis of each of these disorders has to be done separately, which requires more time.

Although there has been a lot of research in the area of ADHD and ASD and their diagnosis using brain scanners and manual observation of subjects for extended periods of time by psychological experts, there has been relatively much less work in the direction of developing automated diagnostic tools for ADHD and ASD using easily available devices (e.g. video cameras). The current methods of diagnosis are not only time consuming but they are also susceptible to human decision making bias. Development of machine learning methods, which can be used as a tool for decision making by human experts, could not only save time but will also help in bringing more objective, repeatable measures into the decision making process.

Current available commercial systems such as QbTest, seek to automate the process of ADHD diagnosis using only the head motion of a person as a proxy for the activity of the subject. The other aspects of head actions including its pose are not taken into account directly. The head motion itself is captured using a normal 2D imaging camera which has limited ability to capture motion in 3D. Facial expression is another aspect which is omitted from current ADHD assessment systems. Facial expressions and gestures can provide important cues about the psychological state of a person. There has been

some work which indicates that facial expressions could be useful in the diagnosis of certain psychological disorders(Wang et al., 2008; Girard et al., 2013). But to the best of our knowledge, until now there has been no research which establishes the relationship between facial expressions and ADHD/ASD.

## 1.1 Motivation and Focus

Although there has been a lot of work in the use of machine learning for object detection tasks there has been considerably less work in harnessing the potential of machine learning and computer vision techniques for human behaviour analysis. Since the face is the most expressive part of a body, it plays a pivotal role in the expression of human behaviour. Therefore, automatic facial expression recognition is one of the key focuses of this thesis. The task of facial expression recognition is particularly difficult as it is affected by multiple factors including shape, appearance and dynamics of facial features, head-pose and expression intensity. It is not clear what is the best way to combine these factors when building accurate models for facial expression recognition. Deep learning based algorithms on the other hand have in recent years shown remarkable performance for object classification tasks. However, these algorithms have yet to show similar performance in the domain of facial expression recognition. Leveraging their full potential might require more optimization and adaptation to the facial expression domain. This thesis proposes a deep learning based approach for automatic facial expression recognition. The proposed approach is adapted to maximize the performance in this domain by combining the use of shape, appearance and dynamics with deep neural networks.

The application of the proposed approach for detecting ADHD and ASD from facial gestures is another key focus of this thesis. The evaluation of any such method requires a database which is suited for this specific task. Therefore, this thesis also presents a database consisting of visual data from clinical and control participants recorded using a RGBD (Colour+Depth) sensor. In addition, a novel approach is proposed to learn and predict ADHD and ASD from visual behavioural data. The facial expression detector proposed in this thesis is employed on this dataset and the output from this detector in combination with other facial gestures is used for automatically predicting ADHD and ASD in the adult population.

The main objectives of this research are as follows:

- To explore the existing neural network architectures and to propose a new deep CNN architecture which is specially adapted for the domain of facial expression recognition.
- To investigate the role of shape, appearance and dynamics in automatic facial expression recognition and propose a way to use them in the framework of deep convolutional neural networks.
- To record a database consisting of visual data from clinical (people diagnosed with ADHD or ASD) and control participants which can be used for evaluating computer vision based algorithms for automatic prediction of ADHD and ASD in adults.
- To propose a novel algorithm which uses facial expression and gestures for automatic prediction of ADHD and ASD in adults.

## 1.2 Research Questions

Deep neural networks have completely revolutionized the field of computer vision due to their high performance on large scale object classification tasks. However, their performance in other computer vision tasks including facial expression recognition remains to be seen. With this we come to our first research question: **Can deep learning algorithms produce a similar level of performance improvement in the field of facial expression recognition, as it has been shown to do in the field of object classification?**

Facial expression recognition requires the analysis of three main aspects of facial features: shape, appearance and the dynamics. Jointly modeling these aspects should (in theory) produce highly accurate models for facial expression recognition. However, the performance of the models depends a lot on the way we model these aspects together. With this we come to our second research question: **How can we model the shape, appearance and dynamics of facial features jointly in a deep learning framework so as to build highly accurate models for facial expression recognition?**

Neurodevelopmental conditions such as ADHD have traditionally been diagnosed by experts through a combination of developmental history, collateral information, psychometrics, behavioural observation and impairment. This procedure not only requires a lot of time and effort, but is also subjective in nature. The subjectivity arises from the fact that observation of patients leaves room for the personal judgement of a clinician. Developing an independent automated recommender system can provide an objective basis for clinical decision making. But such an automated system, which doesn't depend on human expertise, may have its own disadvantages. With this, we come to our third research question: **Can such an automated computer vision based system be implemented which can give accurate prediction comparable to a human expert in this field?**

Facial expressions can provide important cues about the psychological state of a person. There has been some work which indicates that facial expressions could be useful in the diagnosis of certain psychological disorders (Wang et al., 2008; Girard et al., 2013). But till now there has been no work establishing the relationship between facial expressions and ADHD and/or ASD. With this we come to our final question which we aim to answer through this research: **Does facial expressions provide any cues which can help to distinguish between people with and without ADHD/ASD? If yes, how do they relate to existing measures?**

## 1.3 Contributions

This thesis proposes the following contributions to the field of clinical psychology and computer vision:

- Investigate current deep learning algorithms and evaluate them for the task of facial expression recognition.
- A new deep learning framework which uses shape, appearance and dynamics in combination with a deep convolutional neural network, for automatic facial expression recognition.
- A computer vision based approach which uses facial expressions and gestures to make automatic diagnostic predictions about ADHD and ASD. The aim is to help

clinicians in the diagnosis of these neurodevelopmental conditions by making it more efficient and adding more objectivity to their analysis.

- A new database for evaluating computer vision based algorithms on the task of predicting ADHD and ASD diagnosis.
- Establish any possible relationship between facial expressions and the presence of neurodevelopmental conditions like ADHD and ASD in adults.

## 1.4 Publications

- Shashank Jaiswal, Timur Almaev, and Michel Valstar. "Guided unsupervised learning of mode specific models for facial point detection in the wild." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013.
- Shashank Jaiswal, Brais Martinez, and Michel F. Valstar. "Learning to combine local models for Facial Action Unit detection." Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on. Vol. 6. IEEE, 2015.
- Ringeval, Fabien, Björn Schuller, Michel Valstar, Shashank Jaiswal, Erik Marchi, Denis Lalanne, Roddy Cowie, and Maja Pantic. "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data." In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp. 3-8. ACM, 2015.
- Shashank Jaiswal and Michel Valstar. "Deep learning the dynamic appearance and shape of facial action units." 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2016.
- Shashank Jaiswal, Michel F. Valstar, Alinda Gillott and David Daley. "Automatic Detection of ADHD and ASD from Expressive Behaviour in RGBD Data." Automatic Face and Gesture Recognition (FG), 2017 12th IEEE International Conference and Workshops on. IEEE, 2017.

## 1.5 Thesis structure

The overall structure of this thesis is comprised of six chapters. It begins with a brief clinical description of ADHD and ASD in Chapter 2. This chapter gives an overview of the current methods of diagnosis and treatment of ADHD and ASD. It also describes the problems associated with the current methods of diagnosis and briefly reviews some of the recent research works which aim at using automatic behaviour analysis to help in the diagnosis of these conditions.

Chapter 3, gives an overview of the field of automatic face analysis. It describes the core components of this field including face detection, head-pose estimation, facial landmark localization and facial expression recognition and gives a brief review of some of the important methods employed for each of them.

Chapters 4, 5 and 6 represent the main contributions of this thesis. Chapter 4 presents a dynamic deep learning framework for automatic facial Action Unit (AU) detection and intensity estimation. This Chapter describes the main features of this method which includes the architecture of the Convolutional Neural Network (CNN) and its proposed use of local image regions, binary masks and image sequences to jointly learn the appearance, shape and dynamics of each AU. The experimental section of this Chapter describes an extensive set of experiments conducted to evaluate the contribution of different components of the proposed system, its sensitivity to system parameters and compare its performance with other state-of-the-art methods on multiple databases.

In Chapter 5, a new database (KOMAA) is presented which can be used for evaluating methods attempting to use behaviour analysis for automatic detection of ADHD and ASD. The database consists of RGBD (Colour+depth) videos of people who have been diagnosed with either ADHD or ASD or both. It also includes people from the control group who show no symptoms of these conditions. This Chapter describes the collection of this database including recruitment of participants, details of recording scenario and the neuropsychiatric tests which were conducted to gain a deeper insight into the limitations of the current diagnostic tools.

Chapter 6 describes a novel approach which employs the facial AU detection algorithm (described in Chapter 4) along with other 3D face tracking data to encode facial behaviour. This representation of facial behaviour is applied on the KOMAA dataset

(described in Chapter 5) to automatically detect people with ADHD and/or ASD. The Chapter gives a detailed description of this algorithm and the experimental results which demonstrates the potential benefits of using this approach.

The thesis finally concludes in Chapter 7, by summarizing the main chapters and giving directions for future research.

# Chapter 2

# ADHD and ASD: Current Diagnosis, Monitoring and Treatment Methods

Mental health problems are one of the challenging issues facing the field of medical science today. In a recent study in UK, it was estimated that more than a quarter of the population has been diagnosed with at least one mental health disorder (Craig et al., 2015). Mental health conditions can have a profound effect on an individual's life, ranging from short term to lifelong disability. Correct and early diagnosis of such conditions is therefore extremely important in order to provide the right kind of support and help needed. This chapter deals with a special class of such disorders known as neurodevelopmental conditions. Neurodevelopmental conditions effect the development of the brain and usually have their onset in early childhood. These conditions can lead to problems in the social, academic and occupational domain. This chapter describes two of the most common neurodevelopmental conditions: Attention Deficit Hyperactivity Disorder (ADHD) and Autism Spectrum Disorder (ASD). It describes the current diagnosis, monitoring and treatment methods for each of these disorders. It also discusses the clinical co-occurrence (comorbidity) of these disorders and the current challenges in the diagnosis and monitoring of these conditions. Finally, an overview is provided discussing the existing computer vision based approaches which aim to do automatic analysis of such disorders.

## 2.1 Attention Deficit Hyperactivity Disorder

Attention deficit hyperactivity disorder is a neurodevelopmental condition which is characterised by symptoms such as poor attention span, high levels of activity and impulsivity. The prevalence rate of ADHD is estimated to be about 5% (Polanczyk et al., 2007) of the general population, of which half of them (2.5%) are estimated to be adults(Simon et al., 2009). Originally, ADHD was known as a disorder of childhood as the symptoms usually develop at a very early age. However, it has been observed that symptoms often persist into adulthood (Geissler and Lesch, 2011) and the lifelong prevalence of ADHD is now well acknowledged. It negatively affects the quality of life and is found to be associated with low self esteem (Hodgkins et al., 2012) and poor social functioning skills such as increased peer victimization and poor family relationships (Asherson et al., 2012; Mrug et al., 2012).

A lot of research studies have been conducted but the underlying cause of ADHD remains unclear. Some potential risk factors have been identified which are found to be associated with the presence of this disorder. These include biological factors (e.g. genetics, brain structure), environmental factors (e.g. diet, family environment and early neglect) and interaction between inherited and non-inherited factors (Tarver et al., 2014). Genetics is thought to be one of the most significant factors behind this condition. With a high rate of heritability of about 0.7, it is believed that the genes inherited from parents play an important role in developing this condition. However, until now no single candidate gene has been identified to be responsible for this condition. It is believed that interaction between multiple genetic risk variants could be involved in its development (Thapar et al., 2013). Some features of brain structure have also been found to be associated with ADHD. Brain imaging systems like MRI have shown that people with ADHD may have abnormalities like reduced brain volume, abnormal white matter in neural tracts, reduced grey matter in frontostriatal circuits and cortical thinning. Diet is another factor which has been investigated for association with ADHD. Some studies have found children with ADHD to be nutritionally deficient in iron, fatty acids and zinc. However, currently there is not enough evidence to regard these dietary factors as being responsible for development of this condition. Unfavourable family environment, adverse parenting practices and neglect in early development years have also been studied and found to be associated with people having ADHD. The development of this condition is not completely explained by genetics nor do all the individuals who are exposed to the environmental

risk factors go on to develop ADHD. The interaction between genes and environment, where certain genes influences response to external stimuli (environmental risk factors), is believed to play an important role in the development of this condition. This interplay between genes and adverse environmental conditions has been researched in a number of studies e.g. (kah, 2003; K et al., 2006; Nigg et al., 2010). However more research needs to be done to reproduce and ascertain the findings of these studies.

### 2.1.1 Diagnosis

The diagnosis of ADHD is made by clinical experts through a comprehensive evaluation of the patient. Prior to final diagnosis, the patient has to go through a series of assessments which may include diagnostic interview, behavioural observation and scoring of rating scales.

**Clinical assessment:** The assessment process usually starts with an in depth interview between the physician and the patient. The purpose of this being to establish the patient's self perception of his current and childhood ADHD symptoms and impairments. The assessment is centred around the observation of current behavioural symptoms, early (childhood) onset of these symptoms and observation of the symptoms and impairments in multiple environments. The interview can also be done along with family members, partners, carer or friends to take into account their perception of the same as well. Along with behavioural symptoms of ADHD, history of developmental, medical and psychological problems is also taken into account. It is also important to consider the presence of ADHD-like symptoms and other comorbid disorders in the family due to the high rate of inheritance of ADHD.

**Rating Scales and other instruments:** There are number of rating scales available which are often used to to aid the screening and diagnosis of ADHD. The most prominent among them are described below:

- **Adult ADHD Self-Report Scale (ASRS):** This scale consists of a symptom checklist developed by the World Health Organization. The checklist is based on 18 symptoms described in DSM-4 (APA, 2000) criteria for ADHD diagnosis. The DSM-4 criteria was later replaced by DSM-5 (APA, 2013). ASRS can be used as a screening tool for ADHD in adult patients. The check-list consists of two parts:

part A and part B. Part A is the shorter and consists of 6 item which correspond to questions which are found to be most predictive of symptoms associated with ADHD. Part B contains the rest of the 12 items which provide additional cues to help in the diagnosis of ADHD. The patient has to rate each item on the basis of how they felt and conducted themselves in the past 6 months. The whole questionnaire takes around 5 minutes to complete.

- **ADHD rating scale IV (ADHD-RS):** This scale is also a screening measure for ADHD in children. It is also based on DSM-4 criteria and is used to obtain ratings from the parent of the concerned child. There are 18 items in this scale of which 9 items are for measuring inattention and another 9 items for measuring hyperactivity and impulsivity.

Some other screening questionnaires which can be used to guide assessment of ADHD include Brown ADD Scale Diagnostic Form (BADDS) (Brown, 1996) which consists of items to measure attention and executive functioning; and the Wender Utah Rating Scale (WURS) (Ward, 1993) which measures symptoms of ADHD and other related disorders which show comorbidity with ADHD. Conners' Adult ADHD Rating Scale (CAARS) (Conners et al., 1999) is another self-report questionnaire based on the DSM-4 criteria for ADHD. It is available in 2 different versions: one for patients and another for their significant others.

Apart from the screening questionnaires, a number of neuropsychiatric tests have also been developed to measure the cognitive performance of the patients. These can compliment the diagnostic assessment of ADHD. Some popular neuropsychiatric tests related to ADHD are described below:

- **Cambridge Neuropsychological Test Automated Battery (CANTAB):** is computerized test battery aimed at multiple neuropsychological functions including general memory and learning, working memory and executive function, visual memory, verbal memory, attention and reaction time, decision making and response control. This test is more suitable for cross cultural studies because of its use of non-verbal stimuli.

- **Stop signal reaction time task(SSRT):** This task is designed to measure the ability of the brain to suppress a motor response that has already been initiated. It is a

computer based task in which a person is required to respond to a stimuli quickly, except when a stop signal is received. This task is useful for measuring how impulsive a person is, one of the three main symptom of ADHD. A person with ADHD will be slower to respond to the stop signal because of their impulsive nature.

### 2.1.2 Treatment Methods

With the advancement in medical science, it is now possible to effectively treat the core symptoms of ADHD. Treatment methods include medication and psychological therapies. Some important methods of treating symptoms of ADHD are described below:

- **Pharmacotherapy:** Stimulant drugs like methylphenidate and dexamphetamine have been found effective in treating ADHD symptoms and are generally considered the first line of treatment. Atomoxetine is another drug which has been shown to be effective. A number of studies in children and adults have demonstrated the benefits of these drugs (Banaschewski et al., 2006, 2004; Adler et al., 2009). Stimulant drugs have been found to be effective in 70% of cases (Kooij et al., 2010; Biederman et al., 2004; Spencer et al., 2005). The advantage with stimulant drugs is that they are not only effective in treating symptoms of ADHD but also helpful in improving related conditions such as mood swings, low self-esteem, anger, cognitive problems, social and family functions with only mild side effects (Kooij et al., 2010).

- **Psychological therapies:** Among psychological therapies, psycho-education is an important and usually the first step in the treatment of ADHD. It consists of educating the patient and their significant others (parents, partner, friends, etc) about the symptoms and impairments as a result of having ADHD, its prevalence, heritability, treatment options, etc. Providing this information has been found to bring comfort and improve family relationships as a result of having a better understanding of the condition. It also reduces the feeling of guilt and remorse and helps in restoring the patient's social network.

- **Coaching:** is one of the most important psychological therapies available for ADHD. It involves imparting problem solving skills for practical problems typically associated with this condition. ADHD being a life-long condition, prevents

people from learning organizational skills and they are likely to cope poorly and respond inappropriately to problems related to ADHD. The aim of this kind of treatment is to provide structured, supportive therapy by teaching them skills such as efficient time management, acceptance of disorder, dealing with problems in work and relationships among many other things. Cognitive behavioural therapy (CBT) for ADHD is another psychological therapy which aims to impart similar support as a coaching program.

As per current research, the psychological treatment methods such as coaching program and CBT are not recommended as sole method of treatment as they do not aim to relieve the core symptoms of ADHD. The optimal treatment plan follows a multimodal approach which includes a combination of pharmacotherapy, psycho-education, coaching and cognitive behaviour therapy which involves the patient as well as his/her close family, friends or partner.

## 2.2 Autism Spectrum Disorder

Autism spectrum disorder (ASD) is a neurodevelopmental condition which is characterised by symptoms such as difficulty in social communication and restricted, repetitive interests and behaviours. Similar to ADHD, this condition also begins in early childhood. There is increasing evidence that this condition persists in adulthood and is believed to be a lifelong condition. It adversely affects adult functioning including problems in social life, relationships and learning abilities. This leads to a high emotional and financial cost not only for the individual but also their family members. The prevalence rate of ASD is currently estimated to be around 1% in both children (Baird et al., 2006) and the adult population (Brugha et al., 2011). Despite the high prevalence rate (which is increasing every year), people with ASD still find it hard to access diagnostic services.

Genetic factors are believed to play an important role in the development of ASD and it is considered to be one of the most heritable disorders. The heritability in ASD is estimated to be close to 90% (Freitag, 2007). Along with genetic factors, certain environmental factors are also believed to contribute in the development of ASD. These include complications during the prenatal and perinatal stages, complications during birth, exposure to viruses like Rubella and increased paternal age. In a recent study con-

ducted in twin pairs with autism (Hallmayer et al., 2011), it was estimated that prenatal, perinatal, and postnatal environmental factors contribute almost 60%, while the rest is contributed by genetic factors. There is also some evidence that there are differences in the brain chemistry of individuals with ASD. Current research has focussed on the levels of three neurotransmitters: glutamine, gamma aminobutyric acid (GABA) and serotonin. These neurotransmitters are involved in excitatory neurotransmission (glutamine), neurodevelopmental and inhibitory neurotransmission (GABBA), and emotion recognition and response inhibition (serotonin). It has also been observed ASD is much more common in males as compared to females, with a male to female ratio of 4:1 (Fonbonne, 2005).

### 2.2.1 Diagnosis

At present there are no well proven clear biomarkers for detection of ASD, hence the current diagnosis of ASD is based purely on behavioural symptoms. The two most commonly used diagnostic manuals are the International Classification of Diseases, tenth edition (ICD-10) and the Diagnostic and Statistical Manual, fifth edition (DSM-5). According to DSM-5, Autism spectrum disorder is defined as "persistent difficulties with social communication and social interaction, restricted and repetitive patterns of behaviours, activities or interests, present since early childhood, to the extent that these limit and impair everyday functioning". The formal diagnosis of ASD is done by trained clinicians through a combination of clinical interviews, direct behavioural observation and using other diagnostic instruments/screening measures. The interviews can be conducted either with the patient himself (in case of adults) or with parent/caregiver (in case of child) or both. Some commonly used structured interviews and other diagnostic instruments are mentioned below:

- **The Autism Diagnostic Interview - Revised (ADI-R)** (Lord et al., 1994) is a standardized interview which contains 93 items related to behaviour at different ages. It also provides dimensional measures related to communication and language, social interaction, repetitive stereotypical behaviour and age of onset. The interview takes around 2-3 hours and the minimum mental age required for the patient is 24 months.
- **The Diagnostic Interview for Social and Communication Disorders (DISCO)**

(Wing et al., 2002) is another semi-structured interview containing more than 300 questions. The questions deal with social interaction, sterotypical behaviours and impairments related to communication. It is designed to be used with parents and caregivers and there is no age restriction. It takes around two to three hours to complete.

- **Autism Diagnostic Observation Schedule - Generic (ADOS-G)** (Lord et al., 1989) is a semi-structured instrument used for direct behavioural observation. It involves standardized play and communication session which are observed and rated by a trained clinician. There are four different modules available each of which is customized to the patient's level of language skills and cognitive development. Each of the modules takes between thirty to sixty minutes to complete.

- **Childhood Autism Rating Scale (CARS)** (Vaughan, 2011) is a behaviour rating scale for children below the age of six years. It consists of 15 items related to communication (verbal and non-verbal), anxiety, adaptation to change, use of body, imitation, relation to non-human objects, visual responsiveness, auditory responsiveness, near receptor responsiveness, activity level, and intellectual functioning. It involves subjective observation of the child's behaviour which can be completed by clinician, parent or teacher.

- **Autism Spectrum Quotient (AQ)** is a commonly used screening questionnaire consisting of fifty questions related to the symptoms of Autism spectrum disorder. For each question, the subject has to indicate one of the four options: "definitely agree", "slightly agree", "slightly disagree" or "definitely disagree". The questions cover many areas such as social skills, communication skills, imagination, attention to detail and tolerance of change, which are often associated with ASD. In a study (Baron-Cohen et al., 2001b), individuals diagnosed with ASD scored 32 or more in about 80% of the cases as compared to only 2% of the cases in controls. A condensed form of Autism Spectrum Quotient known as AQ-10 (Allison et al., 2012) has also been developed with the aim of having a rapid screening procedure. It consists of ten items selected from the original AQ questionnaire, which show high discriminatory power. In their study (Allison et al., 2012), the authors obtained sensitivity, specificity and predictive values of 0.88, 0.91 and 0.85 respectively, when using a threshold score of 6 in the AQ-10 questionnaire.

### 2.2.2 Treatment Methods

**Pharmacotherapy:** Although the Autism spectrum disorder has been studied for a long time and is estimated to affect a significant part of the population, up till now there has been no medicine which is approved for the treatment of its core symptoms in adults. There is evidence that psychotropic medicines (such as antidepressants, antipsychotics, sleep medication, stimulants etc.) are sometimes used for individuals with ASD, however there is an overall need for proper trials to asses the effectiveness of such medications against ASD symptoms. There is also an urgent need to develop new therapeutic options which can be customized to individual needs.

**Psychological therapies:** The effect of psychological therapies in treatment of Autism spectrum disorder have been studied a number of times. Some studies on Cognitive bahavioural therapy (CBT) has been done (Spain et al., 2015) which have shown a positive effect on the co-morbid mental health symptoms. A modified form of CBT was studied (Russell et al., 2013) for treating OCD and anxiety in adults with ASD. It observed good response on the OCD/anxiety symptoms. The mindfulness technique has also been investigated (Spek et al., 2013) and was associated with significant reduction in anxiety, depression and ruminations. CBT and recreational activity was studied in a controlled trial (Hesselmark et al., 2014) in which people who underwent CBT showed improvement in understanding difficulties and expression of needs. However, no significant difference was observed on the quality of life. In another trial (Eack et al., 2013), Cognitive enhancement therapy was studied in fourteen adults with ASD showing improved cognitive and social skills. In summary, much more research and large scale trials needs to be conducted to find better psychological interventions which are effective not only against the co-morbid symptoms of ASD but also against it's core symptoms.

## 2.3 Comorbidity between ADHD and ASD

It has been observed that the symptoms of ADHD and ASD very often co-occur. People with ASD have been found to have difficulties with attention switching and sustaining attention (Schatz et al., 2002). Problems with attention are so common in people with ASD that it has been suggested to be a part of the cognitive phenotype of ASD (Allen and Courchesne, 2001). Despite the growing evidence of co-occurrence of these disor-

ders, the DSM-4 and ICD-10 guidelines had ASD diagnosis as an exclusion criteria for ADHD. In other words, dual diagnosis of of both ASD and ADHD in the person was not allowed. However, the new DSM-5 manual has recognised the comorbidity between ADHD and ASD and does allow dual diagnosis of these disorders.

The comorbidity between ADHD and ASD has important implications for the impairments, diagnosis and treatment of these conditions. There is some evidence that the severity of psychological problems increases when ASD is comorbid with ADHD (Yerys et al., 2009). In a study (Rao and Landa, 2014) it was observed that children with comorbid ADHD and ASD suffer from greater impairments and face more problems in daily life. Also, the standard treatment for each of these conditions are found to be less responsive for these comorbid conditions (Leitner, 2007).

## 2.4 Problems and Challenges

Although treatment of both ADHD and ASD is a big concern for people (both patients and clinicians) involved in this field, the accurate diagnosis of these conditions still remains a major challenge. As described earlier, current methods of their diagnosis rely on clinical interviews, self-perception and the perception of the family/close friends. This adds subjectivity to the diagnostic procedure and leaves room for human decision making bias. It can lead to erroneous decisions which can prove to be very costly not only in terms of the economics but also the patient's mental health and quality of life.

Another drawback of the current methods of diagnosis is that the behavioural symptoms are either analysed on the basis of patient's self assessment (which may be subjective and unreliable) or by the clinicians through manual observation. Manual observation by the clinicians may be more reliable because of their expertise, however it can be very time consuming, costly and many times not practical. The neuropsychiatric tests currently used to aid diagnostics are designed to measure only cognitive performance. However the neuropsychiatric conditions can alter a person's expressive behaviour which is not measured by such tests. Systems which can automate the task of behaviour analysis can potentially not only save time and money but also help in making the decision making process more objective. Another challenge in the current system is that the patients have to go to clinic for any kind of screening/diagnosis to begin. This prevents early inter-

vention and the delay in treatment can be detrimental to the patient. On the other hand, automatic behaviour analysis systems have the potential to be used in people's homes, schools or workplace to automatically monitor their behaviour and give early warnings when a behaviour typical of a neurodevelopmental condition is detected. This can help in the early diagnosis of such conditions and enable the patients to get help at early stages of their lives. However, development of such automatic systems is difficult because of the limited research in this direction and the unavailability of large databases which can be used for such kind of research. Due to their sensitive nature, visual databases in the field of medicine and mental health are often difficult to collect or be shared with large number of researchers. This hampers research. The following section gives an overview of the early research work which has been done in this area targeted towards automatic behaviour analysis for detection of ADHD and ASD.

## 2.5 Automatic analysis of ADHD, ASD and other related conditions

The use of computer vision techniques for monitoring people for ADHD and ASD is still in its infancy and there has been limited research. Below we describe some of the existing works which aim towards automatic detection of certain behavioural markers which could help in the diagnosis of ADHD and ASD.

### 2.5.1 Detection of ADHD

Some preliminary studies have been conducted to demonstrate the use of depth capturing cameras to monitor the activities of people. For e.g. (HernÃąndez-Vela et al., 2011) uses the method described in (Shotton et al., 2011) to extract a 3D skeletal model of the human body (see Fig.2.1) using RGB-D image sequences. Using this skeletal model, they tracked 14 reference points corresponding to skeletal joints and used them to detect certain body gestures often found in children having ADHD.

For detecting such gestures, they used the Dynamic Time Warping algorithm (Parizeau and Plamondon, 1990). By measuring the similarity between a temporal sequence of images with a reference sequence of a gesture, they demonstrated that they can recognize

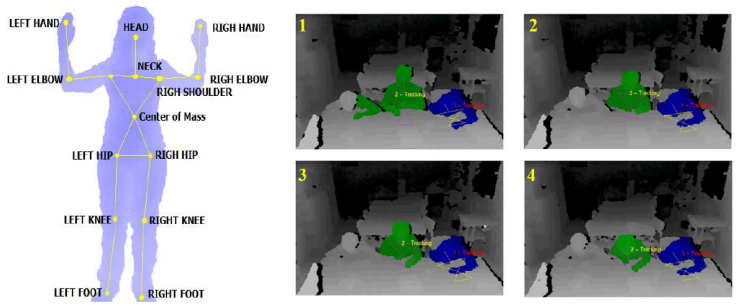


Figure 2.1: The 3D skeletal consisting of 15 distinctive points (left) and a sequence of images where the gesture "lowerhead" is detected (right) (Hernndez-Vela et al., 2011)

a set of defined gestures related to ADHD indicators. For e.g. in Fig. 2.1, they show they are able to detect the gesture "lower head" which is an indicator for ADHD (related to focus of attention).

In (Sivalingam et al., 2012), a system was developed for tracking people across multiple cameras and sensors. They used depth measuring cameras (Microsoft Kinect) to monitor the movement of children in a classroom setting. They used agglomerative hierarchical clustering to segment different objects and tracked different individuals using covariance descriptors. One of the applications they proposed for such a system would be to record the motion tracks and velocity profiles of people, to measure their activity level.

Figure 2.2: Tracking from a single camera (Sivalingam et al., 2012)

QbTest is one of the most successful commercially available systems for monitoring and diagnosis of ADHD. It measures 3 main indicators of ADHD namely, hyperactivity, inattention and impulsivity. It combines head motion tracking with a computer based test. The head motion tracking is designed to measure the hyperactivity of the subject. For this purpose, the subject taking the test is required to wear a head band which has a reflector attached to it. A camera in front of the subject (see Fig. 2.3), tracks the movement of the reflector. However, the system's ability to capture the motion of the

head in full 3D is limited as the camera used does not directly capture the depth of the reflector from the camera. To measure inattention and impulsivity, the subject has to take a computer based test in which he/she has to respond quickly and accurately to certain geometrical shapes displayed on the screen. The whole test lasts for 20 minutes and the head motion is tracked during the entire time. After the test, the result is compared to the norm data corresponding to the subject's age and gender and a report is generated for assessment by clinicians.

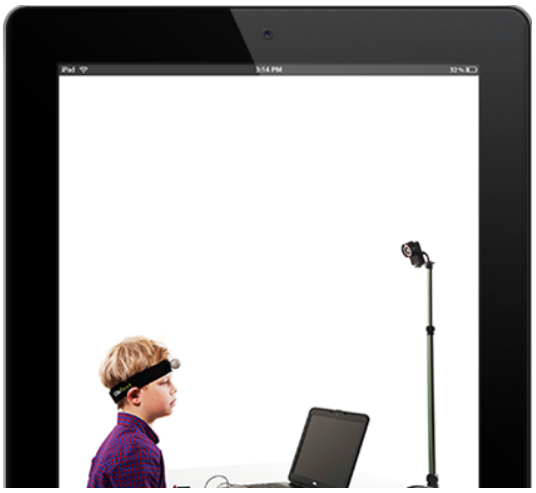

Figure 2.3: QbTest setup.

### 2.5.2 Detection of ASD

One of the pioneering works in the field of ASD diagnosis was done by Hashemi et al. (2012). In this work, the authors developed computer vision based methods to identify certain behavioral markers based on the Autism Observation Scale for Infants (AOSI) which is related to visual attention and motor patterns. For assessing visual attention, they focused on 3 main behavioral markers, namely sharing interest, visual tracking and disengagement of attention. These behavioral markers were detected by estimating the head pose in the left-right direction (yaw) and in the up-down direction (pitch). Head pose was estimated by tracking the position of certain facial features (eyes, nose, ear, etc.). See Fig. 2.4.

Certain motor patterns (specific sequences of muscle movement) have also often been regarded as early biomarkers of ASD. In this work they concentrated on detecting asym-

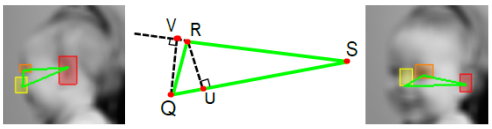


Figure 2.4: Tracking of facial features for estimating the head pose. The triangle created by the left year, left eye and nose is used to estimate the yaw angle (Hashemi et al., 2012)

metric arm position, a motor pattern often found in toddlers diagnosed with ASD (Esposito et al., 2011). 2D pose estimation using an extended Object Cloud Model (OCM)(Miranda et al.) was employed for detecting such asymmetric patterns (See Fig. 2.5).

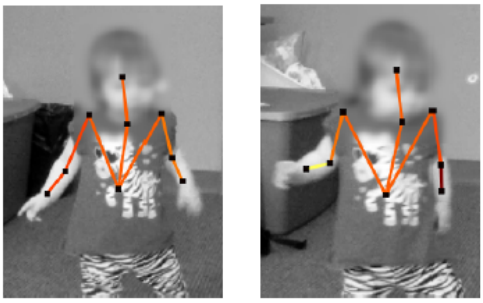

Figure 2.5: Example of symmetric and asymmetric arm with the skeleton automatically placed (Hashemi et al., 2012).

In (Rehg, 2011), the authors presented another computer vision based approach for studying autism by retrieving social games and other forms of social interactions between adults and children from videos. They proposed to do this by defining social games as quasi-periodic spatio-temporal patterns. In order to retrieve such patterns from unstructured videos, the authors represent each frame using a histogram of spatio-temporal words derived from space-time interest points. The frames are clustered based on their histograms to represent the video as a sequence of cluster (keyframes) labels. The quasi-periodic pattern is found by searching for co-occurrences of these keyframe labels in time.

In (Rajagopalan and Goecke, 2014), the authors proposed an algorithm for detecting self-stimulatory behaviour which is a common behavioural marker in individuals with autism. They computed a motion descriptor using dominant motion flow in the tracked body regions, to build a model for detecting self-stimulatory behaviour in videos. Similarly,

in (Rajagopalan et al., 2015), the authors measure children's engagement level in social interactions using low level optical flow based features.

Most of the above mentioned works have concentrated on detecting certain pre-defined behavioural markers which are often associated with either ADHD or ASD in children. However, the performance of these algorithms in actually predicting these conditions by detecting one or a combination of these markers, still remains to be seen. Facial behavioural features is another area which has been ignored by these existing approaches. Faces, which can nowadays be tracked very reliably, can provide important cues for the detection of these conditions through their expressive behaviour.

# Chapter 3

# Automatic Face Analysis

Automatic analysis of human faces has long been a subject of intensive research. This is due its potential applications in various fields such as human-computer interaction, biometrics, psychology, mental health, etc. Automatic face analysis encompasses a wide range of areas which includes detection (localization) of faces in images, facial parts or landmark localization, face recognition (identification), head pose estimation, facial expression recognition, gender recognition, apparent age estimation and many others. This chapter gives an overview of some of the main areas of automatic face analysis relevant to this thesis.

## 3.1 Face detection

Locating the face in a given image is one of the first steps in any kind of facial analysis. Usually it involves outputting a bounding box around each face in a given image. Although current face detection algorithms are quite reliable, it is still considered a challenging open problem due to occlusions, variable lighting condition and head pose.The Viola-Jones algorithm is by far the most popular method for face detection and the first one capable of doing it in real time. It uses haar-like features to train a cascade of Adaboost classifiers. The use of cascaded classifiers makes the algorithm extremely time efficient. Although this algorithm was fast, one of its drawback is that it worked only for near frontal faces.

Another class of algorithms are based on Deformable Parts-based Model (DPM), which can potentially model the deformation between parts (Schneiderman and Kanade, 2004; Mikolajczyk et al., 2004; Zhu and Ramanan, 2012c; Chen et al., 2014). These methods are inspired from or variations of a class of algorithms for generic object detection (Felzenszwalb and Huttenlocher, 2005; Felzenszwalb et al., 2010). A major drawback of these methods is their high computational cost.

Recently deep learning approaches have been demonstrated to show exceptional performance for generic object classification (Krizhevsky et al., 2012a) and detection tasks (Girshick et al., 2014). In (Zhang and Zhang, 2014), the authors proposed a deep CNN to jointly predict the face, pose and locations of facial landmarks. In (Ranjan et al., 2015), the authors employed a DPM approach with features extracted using a deep CNN. Other deep CNN based approaches include (Farfade et al., 2015; Li et al., 2015; Yang et al., 2015). A more detailed overview of the recent approaches to automatic face detection can be found in (Zafeiriou et al., 2015).

## 3.2 Facial landmark detection and Tracking

Facial points are defined as distinctive facial landmarks, such as the corners of the eyes, or the tip of the nose (See Fig. 3.1). Their detection and tracking allows face registration, extraction of geometric features, or local appearance features surrounding the facial points. Most facial point detection and tracking algorithms rely on separate models for the face texture and face shape, which is defined as the set of image locations corresponding to the facial points. Therefore, facial point detection and tracking is posed as the problem of maximising a loss function that depends on the texture model constrained to keeping a valid shape according to the shape model. When non-frontal head poses are considered, they might allow the warping of the facial point locations or of the facial texture to that of a frontal pose. Therefore, it would be possible to perform view-independent AU analysis with models trained just for frontal faces. This is particularly interesting as the daunting cost of manually annotating AUs implies that obtaining labelled examples of the target AUs under a wide range of head poses is practically infeasible.

Face shapes are typically modelled using a statistical shape model (Cootes and Taylor,

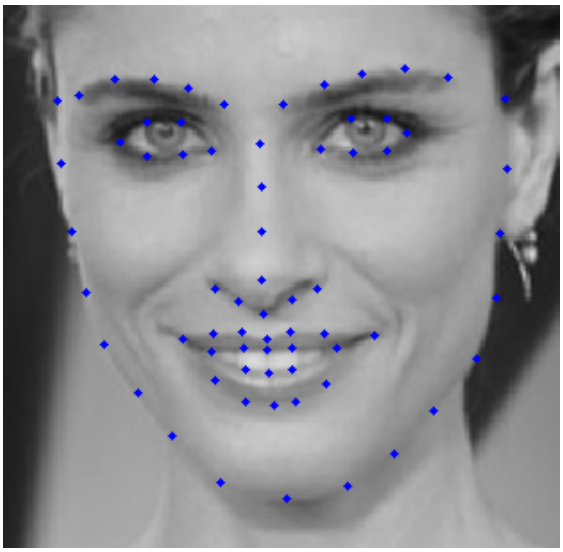

Figure 3.1: Set of 68 facial landmarks.

2004). The possible variations of the face shape depend on two different sets of parameters. Rigid shape transformations are parameterised using a Procrustes transformation, i.e. using in-plane rotation, translation and uniform scaling. Non-rigid transformations are those that cannot be eliminated through Procrustes analysis, and they relate to facial expressions, out-of-plane head rotations and, to some extent, identity.

Other shape models include graphical models, where facial point detection is posed as a problem of minimising the graph energy. For example, (Zhu and Ramanan, 2012a) use a tree to model the relative position between connected points. Here convergence to the global maximum is guaranteed due to the absence of loops in the graph. Similarly, a MRF-based shape model was proposed in (Martinez et al., 2013; Valstar et al., 2010), where the relative angle and length ratio of the segments connecting pairs of points are modelled, making it invariant to both scale and rotation. This shape model is also able to pinpoint incorrect facial point localisations are incorrect, and suggest an alternative that does not affect the other detections. Graph-based shape models are usually more flexible, as the solutions are not restricted to lie in a linear subspace. This does mean that the solution is less constrained, which sometimes leads to larger errors.

When it comes to the modelling of appearance, approaches vary significantly. The most common trends with respect to the way texture information is used include Active Appearance Models (AAMs), Active Shape Models (ASM)/Constrained Local Models (CLMs), and regression-based algorithms.

AAMs (Matthews and Baker, 2004) try to match the whole face appearance with a reference face model. To this end, the facial points are used to define a mesh, and the appearance variations of each triangle within the mesh is modelled using PCA. Face alignment consists of finding the optimal shape and texture parameters so that the reconstruction error is minimised. The appearance models trained for AAMs are often incapable of reconstructing generic faces. Furthermore, the error in the reconstruction is typically measured using the $L_2$ norm, which is not a robust error measure. Therefore, reconstruction errors dominate alignment errors, resulting in poor performance. As a consequence, it is common practise to apply AAMs in person-specific scenarios.

In the ASM framework, the face appearance is represented as a constellation of patches local to the facial points. That is, face locations are represented by extracting a representation over a local patch centred on it. A classifier is trained per point to distinguish between the true target location and surrounding locations. An example of a well-optimised ASM is the work by Milborrow and Nicolls (Milborrow and Nicolls, 2008).

Alternatively, Saragih et al. (Saragih et al., 2009) proposed Constrained Local Models (CLM), where the authors use a non-parametric distribution to approximate the response map. Accordingly, the resulting gradient ascent shape fitting is substituted by a mean-shift algorithm. It is therefore an efficient algorithm that can run in real time. Although the fitting offered is not very precise, it can offer a good trade-off as it can run in real time and offers high robustness. An extension of the CLM was presented in (Asthana et al., 2013), which substitutes the Mean-Shift fitting by a discriminative shape fitting strategy in order to avoid convergence to local maxima.

The work by Zhu and Ramanan (Zhu and Ramanan, 2012b) can be categorised within the ASM/CLM methodology as it uses local appearance models. The authors use a tree-based shape model so that the maximum *a posteriori* likelihood can be attained without using an iterative procedure, and trained a large number of pose-specific experts. This results in a very robust algorithm, capable of performing facial point detection on faces with up to 90 degrees of jaw rotation. However, the precision of the algorithm is often limited and, in particular, it is usually unable to adapt to the presence of expressions.

In regression-based methods the local appearance is analysed by a regressor instead of a classifier. More specifically, given a feature vector, regressors are trained to directly infer the displacement from the test location to the facial point location. Although regression-

based models are very recent, they are one of the dominating trends nowadays and yield the best results to date (Cao et al., 2012; Cootes et al., 2012; Dantone et al., 2012; Martinez et al., 2013; Valstar et al., 2010).

A popular option is to use of random forests regression and fern features to obtain shape estimates (e.g.(Dantone et al., 2012; Cao et al., 2012; Cootes et al., 2012)). This results in very fast algorithms, ideal for low computational cost requirements. Among them, (Dantone et al., 2012) uses conditional random forests to perform regression conditioned to the current face shape. (Cootes et al., 2012) uses random forest voting to generate a response map in combination with the shape alignment strategy of (Saragih et al., 2009). Alternatively, (Valstar et al., 2010) and (Martinez et al., 2013) use Support Vector Regression to obtain point location estimates from stochastically selected local appearance, and aggregate them into a final prediction.

Cascaded regression strategy (Dollár et al., 2010) is currently the most popular for facial landmark localization. Cao et al. (2012) uses random forests in a two-level regression framework and directly regress the full shape, avoiding the shape alignment step. In Xiong and De la Torre Frade (2013), the authors used a cascaded linear regression framework and showed performance comparable to Cao et al. (2012). The primary contribution of their paper was proposing the Supervised Descent Method (SDM) that used Newton optimization for least squares problem to give a mathematical description of the cascaded linear regression framework. SDM has since been extensively studied and extended (Yan et al., 2013a; Zhu et al., 2015; Xiong and De la Torre, 2015).

## 3.3 Head Pose estimation

Estimating the pose of the head can be useful and potentially has a wide range of applications in any kind of facial analysis. For e.g., the pose of the head can be used to estimate attentiveness and the direction of gaze. It can also be used as a gesture to convey agreement or disagreement. It also has implications in learning models for landmark detection and facial expression recognition. Assuming the head to be a rigid object, it has three degrees of freedom (pitch, yaw and roll). A number of approaches have been proposed for estimating the pitch, yaw and roll angles to accurately estimate the pose of the head in 3D space. A brief overview of such approaches is given here.

One popular approach involves the use of appearance templates in which a given test image is matched against a set of template images which have been labelled with pose. The metric used for matching can be Mean-squared error (Niyogi and Freeman, 1996) or cross-correlation (Beymer, 1993). One major drawback of such methods is that they are sensitive to variations in person specific attributes, illumination changes and face localization. In Sherrah et al. (1999), the authors convolved the images with Gabor filters to emphasize certain pose-specific features and to make the matching more invariant to illumination and person specific attributes.

Using an array of face detectors, each trained to detect a specific pose is another widely used method for head pose estimation. In such an approach, given a test image, all the trained detectors are applied one by one and the estimated pose corresponds to the detector with the highest support. Sometimes, a two step procedure is used to make the the process more efficient. For e.g. in Jones and Viola (2003), in order to avoid applying all the detectors to each location of an image, a pose detector is applied first and then depending on its output a face detector specific to that pose is applied. The classifiers used for training such detectors include SVMs (Huang et al., 1998), AdaBoost (Jones and Viola, 2003) and Artificial neural networks (Jones and Viola, 2003).

Another popular approach is to employ non-linear regression methods in order to learn a mapping from the image space to a continuous space of head pose angles. Such methods have the advantage that the output of the trained regressors is in continuous domain as compared to other methods which could only learn certain discreet poses. However, in practice it is a difficult task due to the high dimensionality of the image data. In Li et al. (2000) and Li et al. (2004), the authors use Principal Component Analysis (PCA) to reduce the dimensionality and then Support Vector Regression (SVR) to estimate head pose. Neural networks have been the most popular regression tool for such approaches (Seemann et al., 2004; Stiefelhagen et al., 2002; Stiefelhagen, 2004; Voit et al., 2006, 2008). In Gourier et al. (2004), facial features have been used to train neural networks for estimating head pose. Fanelli et al. (2011) use depth data to learn random regression forests to estimate head pose. Similarly, Kim et al. (2014) use random forests with Binary Pattern (Ojala et al., 1994) based features extracted from random patches. Approaches which employ probabilistic graphical models include Demirkus et al. (2014) and Flohr et al. (2015).

Geometric approaches use the locations of facial features or landmarks to directly esti-

mate the head pose. For e.g. Gee and Cipolla (1994) use the ratio of distances between the locations of outer eye corners, mouth corners and the tip of the nose to determine the head pose. Similarly, Horprasert et al. (1996) use the distances between the inner and outer eye corners and the nose tip to determine the pith, yaw and roll angles. Since only a few facial feature locations are required, these methods are relatively simple and fast. However, they are sensitive to errors in the localization of facial features and require precise locations of facial landmarks to give reliable results.

As mentioned earlier, the high dimensionality of facial images implies that apart from pose information, they also contain information about identity, facial expressions, etc. A common technique used is to project the images onto lower a dimensional manifold in order to reduce other modes of variation. However, the challenge lies in finding a dimensionality reduction technique which only recovers the variation in pose while ignoring other modes of variation like identity and facial expressions. PCA and its kernel version (KPCA) are common techniques used for unsupervised dimensionality reduction (McKenna and Gong, 1998). Other manifold embedding techniques include Isometric feature mapping (Raytchev et al., 2004; Tenenbaum et al., 2000), Locally Linear Embedding (Roweis and Saul, 2000) and Laplacian Eigenmaps (Belkin and Niyogi, 2003). However, since these are unsupervised techniques there is no control which modes of variations are selected. In order to resolve this problem, Srinivasan and Boyer (2002) computed pose specific eigen spaces by using PCA on groups of images which share the same discreet head poses. A pose biased distance metric was also proposed (Balasubramanian et al., 2007) to manifold embedding for head pose estimation. A detailed survey of various head-pose estimation methods can be found in (Murphy-Chutorian and Trivedi, 2009).

## 3.4 Facial expression recognition

Facial expressions are an essential component of the way humans communicate. Due to evolutionary processes humans are able to easily recognise each other's facial expressions. However, automatic recognition of facial expressions is still an unsolved problem despite many years of research. The problem of facial expression recognition is currently posed as either the recognition of facial muscle actions (Action Units) defined according to the Facial Action Coding System (FACS) (see Fig.3.2), or the emotions expressed by

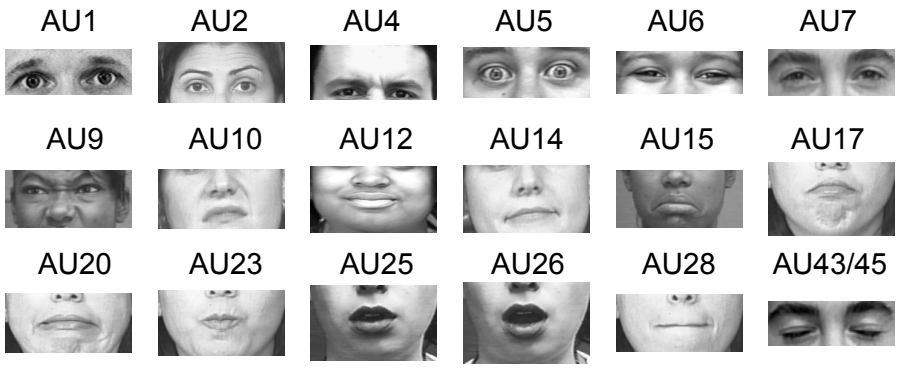


Figure 3.2: Facial Action Units defined according to Facial Action Coding System (FACS). Images taken from *https://www.cs.cmu.edu*

the facial muscle actions (affect analysis). The latter has been researched in two fundamentally different ways. According to one viewpoint, there are discreet categories of basic emotions which arise from separate neural pathways. It is based on the observation made by Ekman and his colleagues that humans are hard-wired to produce and interpret certain expressions which convey universally recognized emotions. These emotions are usually divided into 6 basic classes: Anger, Disgust, Fear, Happiness, Sadness and Surprise (see Fig. 3.3). According to another viewpoint, emotions can be expressed as a point in the space formed by a few affect dimensions. One of the most popular dimensional model of emotions is the Circumplex model developed by Ressel (1980). In this model the emotions are assumed to lie in a two dimensional circular space formed by valence and arousal. Valence represents positivity (pleasant) or negativity (unpleasant) of emotion, while arousal represents calming or exciting emotional states. This thesis focusses on facial expression analysis by recognition of facial muscle actions (Action Units). A brief overview of the important approaches used for automatic facial expression recognition, is given below.

### 3.4.1 Traditional approaches

Traditional approaches to automatic facial expression recognition followed a two step procedure involving feature extraction and recognition using standard machine learning techniques. Such approaches can be broadly classified according to different types of features extracted (e.g. geometric, appearance, etc.) and according to the type of machine learning method used (e.g. SVM, Decision trees, etc.). Below we describe each category of these methods and the important works related to them.

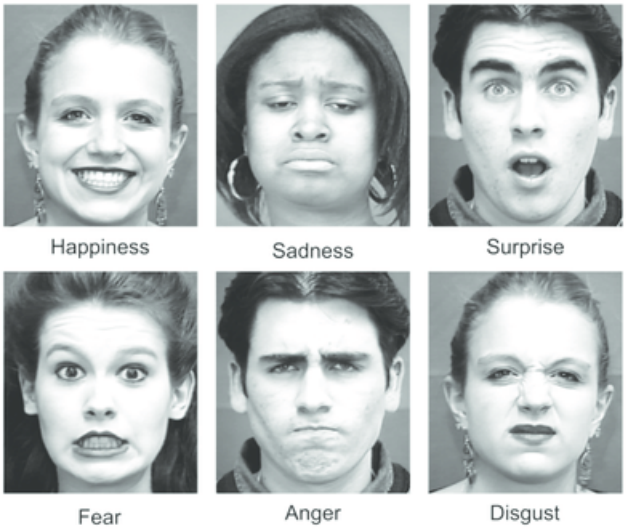


Figure 3.3: Photographs describing six basic emotions from the Cohn-Kanade database (Lucey et al., 2010).

**Appearance features:** Appearance features encode texture information useful for representing facial deformations like wrinkles, furrows and bulges. A number of approaches exist to extract appearance based features. Local Binary Patterns (LBP) is one of the most popular descriptors used for appearance modelling in facial expression recognition (Zhao and Pietikainen, 2007; Shan et al., 2009). LBP features are extracted by comparing the intensity at each pixel with other pixels in its local neighbourhood and computing a sequence of binary digits known as LBP codes (Ojala et al., 1994). Histograms of these LBP codes are usually computed over a spatial region to form the final feature vector. The advantage of LBP features is that they are simple to compute and highly invariant to changes in illumination. Gradient based descriptors like Histogram of Oriented Gradients (HOG) (Dahmane and Meunier, 2011; Dhall et al., 2011) are another class of feature descriptors often used for automatic facial expression recognition. HOG features are computed by counting the discreet gradient orientations in an image region, usually a small rectangular cell. Scale Invariant Feature Transform (SIFT) is another feature descriptor similar to HOG, computed from gradient orientations (Ren and Huang, 2015; Soyel and Demirel, 2010).

Representation using Gabor filters is another popular method for facial expression recognition (Littlewort et al., 2011; Wu et al., 2011b; Glodek et al., 2011). In this method, a set of Gabor filters each of different scales and orientation, are convolved with the facial images. Gabor filters can be sensitive to wave-like structures and hence can easily

represent facial features like wrinkles and bulges, etc. (Shan, 2008). Their differential nature provides robustness to illumination changes and the smoothness makes them robust to misalignment of faces. However, computing convolutions with a set of Gabor filters is not only computationally expensive, it also results in a very high dimensional representation. Dimensionality reduction techniques such as PCA are often used to remove redundant features and keep only the relevant ones. Gabor features have also been combined with LBP in the form of Local Gabor Binary Pattern (LGBP) features (Zhang et al., 2005). LGBP features are computed by first convolving with a bank of Gabor filters and then computing LBP codes on top of the convolved images. Histograms are computed for each of the images which are concatenated to form to a single feature vector. These features have been shown to outperform LBP features for facial expression recognition (Moore and Bowden, 2011). Other appearance descriptors which have been used for facial expression recognition include the Discreet Cosine Transform (DCT) (Hesse et al., 2012; Kaltwang et al., 2012), Haar-like filters Whitehill and Omlin (2006); Kim and Pavlovic (2010) and quantized local Zernike moments (QLZM) (Sariyanidi et al., 2013).

Apart from the type of features extracted another very important aspect to consider is where (or how) these features are extracted from the facial images. Different strategies have been proposed in the literature. The two most common approaches are: the so-called holistic and part-based approaches. The holistic approach represents the appearance of the full face bounding box and extracts features directly from it. The bounding box is usually split into a number of rectangular blocks to encode spatial information (Jiang et al., 2011). Alternatively, the face appearance can be represented by a combination of local image patches around each facial landmark (Zhu et al., 2011). In both cases, a set of feature vectors are extracted from different image patches (be it subdivisions of the bounding box or patches around landmarks). These vectors are concatenated into a single feature vector, which is the input to the training and inference routines. Holistic methods are capable of representing the whole face appearance and not only patches around points. Furthermore, they are sensitive to flexible shape deformations (e.g. the lip stretching which is associated with a smile). However, they offer poor registration, in the sense that since face images are only globally aligned, each pixel will refer to slightly different parts of the face on different examples. Thus, each feature encoding appearance will refer to a slightly different part of the face, and extracting generalising fine-grained patterns from holistically-computed appearance feature vectors is challeng-

ing. Part-based models offer instead a much better registration. However, they are less sensitive to flexible movements, and they do not represent the whole face. Furthermore, works such as Lucey et al. (2011), take the registration step to an extreme where all faces are registered with a piecewise affine transformation into neutral frontal pose (thus maximising registration and yet eliminating an important amount of the expressive information). Remarkably, this strategy offers good performance, and highlights the paramount importance of a good face registration strategy (Jiang et al., 2014a). However, whether there is a better intermediate option in which less expressive information is lost, is a reasonable question.

The work by Jiang et al. (2014a) offered an alternative solution. It consists of using the facial landmarks to create a mesh, as in typical works on active appearance models Matthews and Baker (2004). Then, face regions are defined by merging some of these triangles. Which triangles to fuse is manually defined, but the decisions are based on domain knowledge relating different regions to the facial muscles responsible for the AU. This strategy showed superior performance in combination with different feature extraction approaches and for two state-of-the-art databases.

**Shape features:** Shape features encode information about the geometry and relative locations of facial features usually represented by facial landmarks. The typical steps involved in the extraction of shape features include detection of facial landmarks (see Section 3.2) and use of the location of detected landmarks to compute geometric features. The most simple and commonly used approach is to directly use the x and y coordinates of facial landmarks concatenated together (Rudovic et al., 2012; Lucey et al., 2007). Pantic and Rothkrantz (2004) used differences in the location of facial landmarks between the current frame and a frame showing a neutral face. Similarly, difference between facial landmark locations and distances (from the first frame in a video sequence) were used in Pantic and Patras (2006). Apart from these features, Valstar and Pantic (2012b) employed additional features such as the angle that a line connecting a pair of landmarks makes with the horizontal axis, speed of the facial landmark displacements and the coefficients of second-order polynomial fitted to describe the motion of facial landmarks within a 7 frame temporal window. Valstar et al. (2015c) used a set of shape features consisting of displacement of facial landmarks from the mean shape and from the previous frame, and distance of each facial landmark from the median locations of stable (invariant to expressions) facial landmarks. In this work, the facial landmarks

were also split into groups corresponding to eyes, brow and mouth region and within each group, the Euclidean distances between two consecutive facial landmarks and the angle between two consecutive line segments formed by joining two consecutive landmarks were computed and formed a part of the geometric feature vector.

Shape features are invariant to changes in illumination, however illumination variation can affect the accuracy of the facial landmark detection and hence can indirectly affect the shape features. They also usually have a low dimensionality and simple to compute. However, such features cannot model appearance changes which can be important for recognition of certain facial expressions or AUs. Approaches which have used a combination of appearance and shape features have usually outperformed others (Nicolle et al., 2012; Senechal et al., 2011).

**Spatio-temporal features:** Spatio-temporal features aim to encode temporal information within a range of frames in a temporal window, along with spatial features. Temporal information can be used to model the temporal structure of facial expressions which can compliment the spatial features in facial expression recognition. It is particularly useful for AUs or expression which appear similar in the spatial domain but display a significant difference in their dynamics. A primary example for such a case is AU43 (eyes closed) and AU45 (blinks) (Sariyanidi et al., 2015). Temporal information is also important in learning models for temporal segmentation of facial expressions (onset, peak, offset, etc.) and other high level tasks such as distinguishing posed expressions from spontaneous ones.

A number of approaches have been proposed on spatio-temporal features. Among the appearance based spatio-temporal features, TOP features extracted from three orthogonal planes: the spatial plane (X-Y), and two spatial-temporal planes (Y-t) and (X-t), where X, Y and t represent the horizontal, vertical and time axis respectively. TOP features were first proposed by Zhao and Pietikainen (2007) in the form of LBP-TOP where they proposed extracting LBP features from the three orthogonal planes for recognition of 6 basic emotions. Their proposed features have also been applied for facial AU detection (Jiang et al., 2014b). Other variations of TOP features have also been proposed such as LPQ-TOP (Jiang et al., 2011, 2014b) and LGBP-TOP(Almaev and Valstar, 2013). Due to their ability to encode temporal information in addition to spatial features, TOP features have been shown to outperform their static counterparts. However, these features are known be sensitive to face alignment errors (Sariyanidi et al., 2015). Other draw-

backs include high computational cost and high dimensionality of the feature vectors produced.

Other alternative approaches for encoding spatio-temporal information include dynamic Haar features (Yang et al., 2007), convolution with spatio-temporal extensions of Gabor filters (Wu et al., 2010) and Independent Component filtering (Long et al., 2012). To encode temporal information, Geometric feature based approaches typically use differences in the location of facial landmarks between the current and the neutral frame Pantic and Rothkrantz (2004); Baltrusaitis et al. (2015) and the speed of displacements of facial landmarks Valstar et al. (2015c).

**Learning and Inference:** In traditional approaches, the features extracted from the facial images are used in learning a machine learning based classifier (for occurrence detection) or a regression model (for facial expression intensity estimation). Various approaches have been used for learning the statistical models for facial expression recognition. Support Vector machines (SVM) have been by far the most popular technique employed for learning classification models for facial expression recognition (Zhong et al., 2012; Long et al., 2012; Jiang et al., 2014a; Valstar et al., 2015c). Other methods include Artificial Neural Networks (ANN) (Padgett and Cottrell, 1997; Jaiswal et al., 2015), k-Nearest Neighbour (kNN) (Lyons et al., 1999; Donato et al., 1999), Relevance Vector Machines (Datcu and Rothkrantz, 2005) and Ensemble learning methods such as Random Forests (Pfister et al., 2011; El Meguid and Levine, 2014), AdaBoost (Yang et al., 2011) and GentleBoost (Hamm et al., 2011). These methods take features from an image or from a sequence of images (spatio-temporal features) to learn and predict facial expressions. If temporal information is modelled, it is modelled at the feature level. However, another class of methods model the temporal information at the classifier/regressor level. Such methods usually employ graphical probabilistic models such as Hidden Markov Models (HMM) (Lien, 1998; Schmidt et al., 2010) and Conditional Random Fields (CRF) (Chang et al., 2009; Walecki et al., 2015). HMM and CRF are the linear chain variations of the more general Dynamic Bayesian Networks (DBN) (El Kaliouby and Robinson, 2005; Tong et al., 2010). Graphical models like DBNs are capable of encoding dependencies between a set of random variables which evolve in time. Hence they can not only model temporal information but they are also capable of modelling the relationships between different facial expressions. Apart from these methods, some other approaches which model facial expression dynamics include the work by Valstar

and Pantic (2007) which use a combination SVM and HMM for facial AU recognition and the work by Nicolaou et al. (2012) in which the authors proposed an extension of RVM capable of learning temporal information for dimensional and continuous emotion prediction.

A few other approaches have focussed on alternative ways to boost the performance of facial expression recognition algorithms. Chu et al. (2013b) proposed the Selective Transfer Machine (STM) for personalized facial AU detection. To overcome the inability of a generic classifier in generalizing to previously unseen subjects (due to differences in facial morphology), STM learns a personalized classifier by re-weighing the training examples according to their relevance with the test subject. Similarly, Almaev et al. (2015) proposed a method for person specific AU detection by learning the latent relations between subjects using a reference AU which is easy to annotate, and then transfer this information for the learning to detect other AUs. Learning the statistical relationship between different facial expressions is another area which has been exploited to boost the performance of facial expression recognition algorithms. For e.g. AU6 and AU12 often co-occur together while AU28 and AU12 do not. Such co-occurrence statistics can help in improving the accuracy of the detection algorithms. To model such dependencies, a number of methods have been exploited which includes Restricted Boltzmann machines (RBM) (Wang et al., 2013), DBN (Tong et al., 2010), Discriminant Laplacian Embedding (Yüce et al., 2015) and a multi-task extension of multi kernel SVMs (Zhang et al., 2014a).

**Approaches for Micro-expression recognition:** Micro expressions are a sub-class of facial expressions characterised by very short duration and low intensity changes in the shape/appearance of facial regions. The duration of such facial expressions range from $1/3$ to $1/25$ of a second (Matsumoto and Hwang, 2011; Yan et al., 2013b). These subtle expressions are usually involuntary and therefore can encode emotions that people may not want to display. Recognition of such emotions can have valuable applications in the field of psychology and forensics. Automatic recognition of such expressions is extremely challenging due to the very nature of such expressions (short duration and low intensity). In recent years a number of approaches have been proposed for micro-expression recognition. Most approaches employ methodologies which are similar to the ones used for AU detection and basic emotion recognition. Polikovsky et al. (2009) proposed 3D gradient descriptors for recognition of micro-expressions. They evaluated their

method for recognising 13 different micro-expressions in a database which was recorded using high speed camera at 200 frames per second. Wu et al. (2011a) used Gabor features and a combination of Gentle-Boost and SVM for micro-expression recognition. Their system was evaluated on METT database (Ekman, 2004).

Both the above methods were evaluated on databases containing posed micro-expressions. However, in recent years a number of databases have become available which are labelled with spontaneous micro-expressions. SMIC(Li et al., 2013) and CASME II(Yan et al., 2014) are examples of such databases. Pfister et al. (2011) proposed one of the first method which was evaluated for spontaneous micro-expressions. They used a temporal interpolation model to artificially increase the number of frames corresponding to micro-expressions. By increasing the number of frames, the authors claim they were able to achieve statistically more stable feature extraction results. The interpolated frames were used for extracting LBP-TOP features followed by classifier training using multiple kernel learning. Li et al. (2017) employ feature difference analysis for micro-expression detection. For classification of the detected micro-expression frame sequence, the authors employ Eulerian magnification method for magnifying subtle motions in the videos. They also use temporal interpolation model to overcome the difficulty posed by short duration of micro-expressions. This was followed by feature (HOG and LBP) extraction from three orthogonal planes (formed by 2 spatial and 1 temporal axis). These features were used for learning a linear SVM. The method was evaluated on both SMIC and CASME II databases and achieved an accuracy of 75% and 78% on these databases respectively. Another notable work in this direction is by Sariyanidi et al. (2017) in which the authors proposed to express facial expression variation as a linear combination of localised basis functions. The coefficients of these basis functions represent intensities of facial muscle movement. The proposed method achieved an accuracy of 65% on SMIC database.

### 3.4.2 Deep learning approaches

In the past few years, deep learning approaches have become extremely popular in the field of computer vision. This is due to their success in tackling different computer vision problems, achieving state-of-the-art performance in object classification and detection (Krizhevsky et al., 2012a; Girshick et al., 2014; Simonyan and Zisserman, 2014; He

et al., 2016), image segmentation (Shelhamer et al., 2016), face recognition Parkhi et al. (2015), etc. Inspired by their success, a number of works have employed deep learning for facial expression recognition recognition. In contrast to traditional approaches, deep learning approaches do not a have separate feature extraction and model learning step. Instead, features are learnt in a hierarchical manner through the multiple layers of an artificial neural network (ANN). Convolutional Neural Networks (CNN) are by far the most widely used type of ANN. These are characterized by their shared weight architecture and local connectivity between neurons. Below we describe some of the important approaches which use a CNN based framework for facial expression recognition. Here we divide such approaches into two categories: CNN models of basic emotions and CNN models of facial Action Units.

**CNN models of basic emotions:** Fasel (2002) was one of the first to use CNNs for the task of facial expression recognition. He used a 6 layer CNN architecture (2 convolutional layers, 2 sub-sampling layers and 2 fully connected layers) for feeding face images of size 64x64 pixels. He experimented with 2 versions of his architecture. In the first version, the filter size in the first convolutional layer was fixed to 5x5 pixels, i.e. the features at the first layer were extracted at a single scale. In the second version, the features at the first layer were extracted at multiple scales, using filters of 3 different sizes (5x5, 7x7 and 9x9 pixels). This CNN consisted of 3 different streams corresponding to each scale which are connected to each other only at the fully connected part of the network. Fig. 3.4 shows the architecture of this network. The advantage of this network was that it was able to extract features at multiple scales. This can be useful because different facial expressions occur at different scales. However, the activation functions used in their networks were sigmoidal which are susceptible to vanishing gradient problem, especially in case of deep networks. Additionally, no temporal or explicit shape information was used in learning the models for facial expressions.

Rifai et al. (2012) proposed a method to disentangle features which are discriminative for facial expressions from all other features. They use features from a CNN (1 layered) whose filters were pretrained using Contractive Autoencoders (CAE). The feature output from the CNN provided input for a semi-supervised feature learning framework called Contractive Discriminative Analysis (CDA). CDA is a semi-supervised version of CAE, in which the input is mapped onto 2 distinct blocks of features. One of the blocks learns features which are discriminative for facial expressions, while the other block learns all

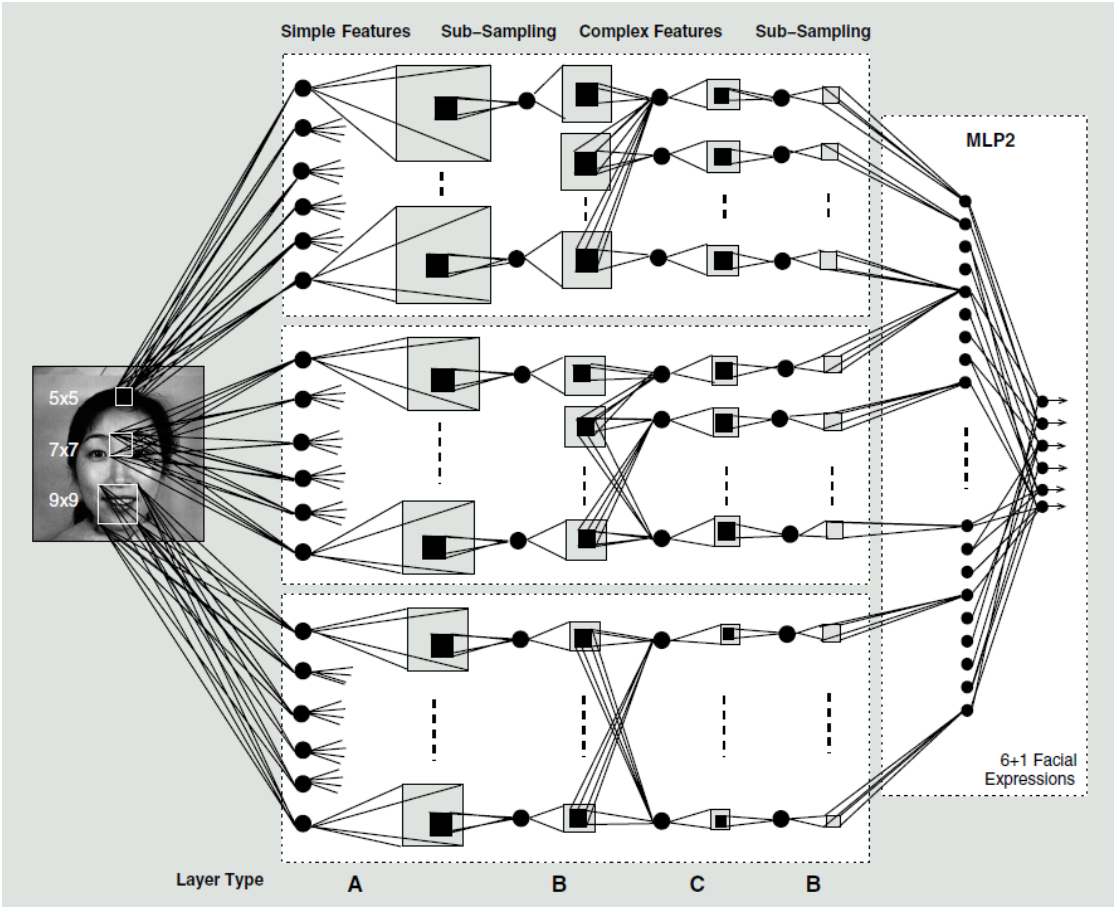


Figure 3.4: CNN architecture consisting of 3 streams used by Fasel (2002)

other features (see Fig. 3.5). Both the blocks are learnt so as to jointly reconstruct the input. The discriminative features from the first block are then used to learn a SVM for facial expression classification. The main advantage of this approach is that it is able to separate out the features useful for facial expression recognition (invariant to pose, identity, etc.) while the pose and identity specific information gets filtered out through the other block. However, the CNN used consisted of only one layer which prevents learning of rich hierarchical features. The CNN weights were also pre-trained and were not learnt in end-to-end manner. Additionally, no dynamic or shape information was utilized.

Liu et al. (2015), used 3D CNN for dynamic learning of facial expressions in a video. They proposed a CNN architecture which can jointly localise certain dynamic parts of a face ("action parts") and encode them for facial expression classification. Their architecture consisted of 7 layers as shown in Fig. 3.6. The input to the network consists of $n$ contiguous frames of a video in which the face has been detected and normalized to size

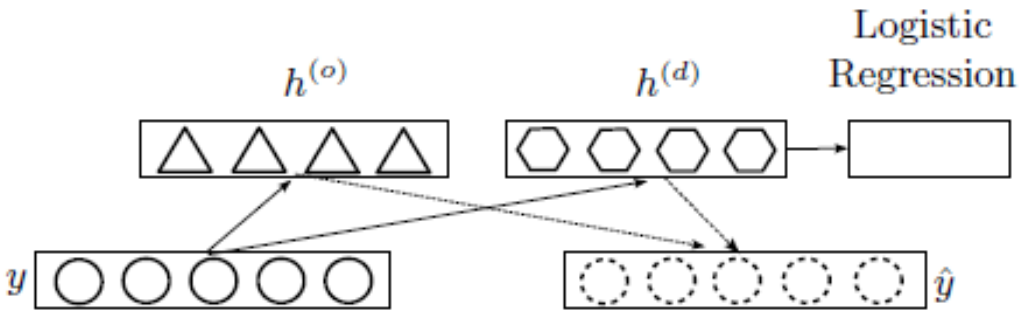


Figure 3.5: Contractive Discriminative Analysis (CDA) framework used by Rifai et al. (2012).

of 64x64 pixels. The first layer is a convolutional layer consisting of 64 spatio-temporal filters of size 9x9x3. The resulting feature maps are mean-pooled with non-overlapping blocks of size 2x2x1. The resulting feature maps are then convolved with a bank of class (of facial expression) specific part filters to compute part detection maps. These part detection maps and a set of deformation maps are summed together (with learned weights), to enforce spatial constraints on the part detections. The resulting feature maps are partially connected to the sixth layer. In this layer, all the feature maps corresponding to a particular part are fully connected to a separate set of nodes whose size is the same as that of the number of facial expression classes. This layer is fully connected to the last layer which outputs decision values for each of the facial expression classes. The advantages of this approach include implicit detection of facial parts useful for facial expression recognition and learning temporal information by using a sequence of contiguous frames as input to a 3D CNN. However, no explicit shape information was used to model any facial part. Moreover, the temporal information which can be learnt is limited only to $n$ consecutive frames. Longer range temporal information (greater than $n$ frames) which can be useful for certain facial expressions, are ignored by this approach.

Jung et al. (2015) used a deep CNN to learn temporal appearance features for learning facial expressions. Additionally, they also employed a deep neural network to learn temporal geometric features from detected facial landmarks. Both the networks were learned independently to predict facial expression. The output decision values from each of the networks were combined (linear combination) to compute the final score for any example face image. The CNN architecture used for learning appearance features is shown in Fig. 3.7. It consists of 2 convolutional layers and 2 fully connected layers. Each convolu-

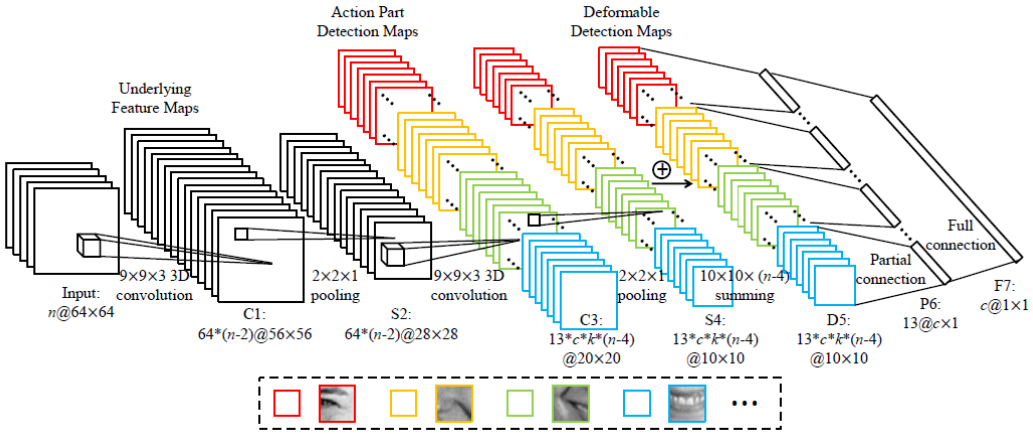


Figure 3.6: CNN architecture used by Liu et al. (2015).

tional layer is followed by a max pooling layer. In order to encode appearance dynamics, the network takes in as input, a sequence of face images re-sized to 64x64 pixels. The deep network for learning temporal geometric features consists of 2 hidden layers (fully connected) as shown in Fig. 3.7. The input to this network consists of coordinates of facial landmarks tracked in a sequence of face images. The advantages of this approach is the explicit modelling of temporal shape and appearance. The main drawback is that the shape and appearance information is not modelled jointly in the network. Due to this, it may not be able to learn the best combination of shape and appearance parameters. As with previous works, this approach is also limited in the amount of temporal information which can be learnt as only a fixed number of frames can be used as input to the CNN.

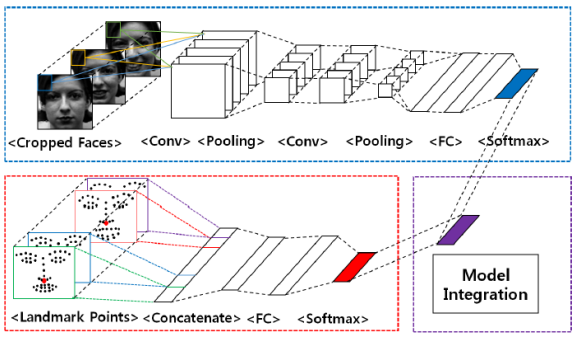


Figure 3.7: CNN architecture used by Jung et al. (2015).

Another notable work in this area belongs to Kahou et al. (2013), in which the authors combine face models learnt using a deep CNN architecture with various other models for facial expression classification. They combined a CNN face model with a bag of words

model for mouth region, a deep belief network for audio information and deep autoencoder for modelling spatio-temporal information. A weighted average of the predictions from all the models, was used for classifying the emotion expressed by the primary human subject present in short video clips. The CNN architecture used for modelling the face is shown in Fig. 3.8. The input to this network consists of face images of size 40x40 pixels which were cropped (randomly during training) from images of size 48x48 pixels. The network consists of 3 convolutional layers and 1 fully connected layer. Each convolutional layer is followed by either a max pooling or average pooling layer. The network also consists of 2 local response normalization layers after the first and second pooling layers. The output layer (softmax) has 7 units corresponding to the 7 basic emotions. The main advantage of this method was the fusion of different modalities corresponding to appearance, dynamics and audio, resulting in robust models for facial expression recognition. However, no explicit shape information was utilized.

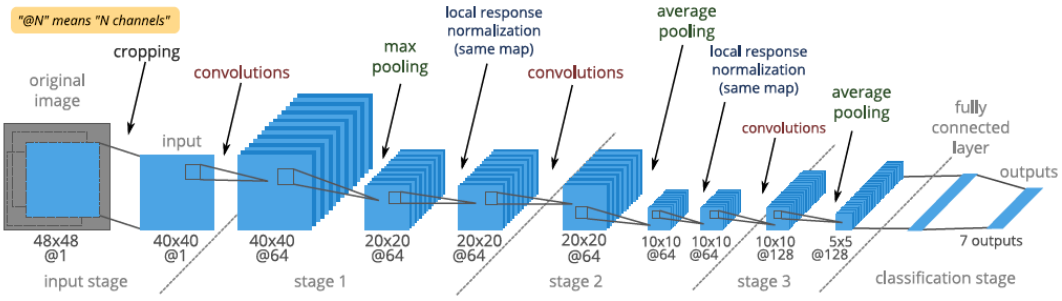


Figure 3.8: CNN architecture used by Kahou et al. (2013).

Some other recent works include the Curriculum Learning approach by Gui et al. (2017) in which the training samples are sorted into a sequence of subsets. These sequence of subsets are formed so that easier samples occur first and the samples become gradually harder as the sequence continues. The authors show that reordering the training samples in such a way benefits optimization of the facial expression models where a model learns to recognize easier samples first followed by harder samples. Another interesting work was proposed by Kosti et al. (2017) in which a CNN is used to learn context information for modelling human emotion . In this work, the authors proposed a CNN consisting of 2 modules: one which learns features from the image region containing the human subject and another which learns global features from the entire image (including background) for providing context support. The approach was evaluated on a new EMOTIC database showing the advantages of using context for emotion modelling.

**CNN models of facial Action Units:** Relatively fewer CNN based approaches have been proposed which are specialized for targeting the problem of facial AU detection. Gudi et al. (2015) used a deep CNN consisting of 3 convolutional layers, 1 sub-sampling layer and 1 fully connected layer to predict the occurrence and intensity of Facial AUs. This approach showed performance comparable to the performance of other participants of the FERA-2015 Valstar et al. (2015c) challenge. A similar architecture was used by Ghosh et al. (2015) in which they evaluated their approach for cross dataset performance using BP4D, DISFA and CK+ databases. A region learning method was proposed by Zhao et al. (2016) to learn local appearance changes due to different facial AUs. They employed a "region layer" whose weights are shared only within a small local patch of the face defined by the cells of a uniform rectangular grid of size $8 \times 8$.

The work by Tősér et al. (2016) focussed on augmenting the training data to increase the amount of head pose variation. They used the 3D face models provided in the BP4D dataset to create new head orientations from the existing images. Their evaluation also showed that training a multi-label CNN for learning all AUs together in a single CNN, is much harder and showed inferior performance compared to a single label one. In another interesting work by Han et al. (2016) proposed to integrate boosting into CNNs, by introducing an incremental boosting layer. The approach works by selecting and updating highly discriminative neurons at the fully-connected layer to incrementally learn a strong classifier over time.

**Relation to our work:** As we learnt in Section 3.4.1 that the traditional approaches to facial expression recognition were based on extracting the three main kinds of facial features: shape, appearance and dynamics. These features compliment each other and are known to be important for learning robust models of facial expressions. Deep CNN models on the other hand have proved to be very successful and are fast replacing the traditional hand crafted feature approaches due to their ability to learn features (from the data) which are specialized for the target problem. However, most of the existing deep CNN based approaches (Fasel, 2002; Gudi et al., 2015; Liu et al., 2015), do not utilize all the key facial features i.e. its shape, appearance and dynamics. Those which do utilize all three of them (Jung et al., 2015), do not learn them jointly. We believe that in order to find the most optimum combination of these features, it is necessary to model them jointly. Also, almost all the above CNN based approaches use a fixed time window to learn the temporal information. This limits access to temporal information within

a specific time window only. On the other hand, different facial AUs can occur over different time scales. For e.g., Smile (AU12) can last for much longer time as compared to blinks (AU45) which occurs only for a short interval of time. Even within a specific AU, the variation in its duration of occurrence can be quite large.

The approach for recognition of facial expressions presented in this thesis is based on the deep CNN framework. However, in contrast to existing methods, the method proposed in this thesis uses all key facial features (shape, appearance and dynamics) and attempts to jointly learn them in a single CNN. Additionally, it also attempts to overcome the problem of a fixed time window by employing Bi-directional Long Short Term Memory (BLSTM) neural networks for learning long term temporal dependencies.

# Chapter 4

# Dynamic Deep Learning Facial Expressions

Automatic facial expression recognition in terms of facial Action Units (AUs), is a challenging problem primarily due to the high degree of variation in the visual appearance of human faces (caused by person specific attributes, multiple poses, etc.), low intensity activation of spontaneous expressions, non-additive effects of co-occurring AUs and the scarcity of training data. Accurate facial AU recognition involves the analysis of three main facial features: its shape, appearance and dynamics. Each of these can be considered a source of complimentary information for the modelling of facial action unit detectors. We hypothesize that learning all the three features jointly can produce highly accurate models for facial AU recognition. However, the performance of the models depends a lot on how one models these features together.

This chapter describes a deep learning based framework for facial AU detection in sequence of images (videos). In particular Convolutional Neural Networks (CNNs) are used to model the appearance, shape and the short-term dynamics of facial regions for AU detection. In contrast to previous approaches, this system learns all the key features (appearance, shape and dynamics) jointly in a deep CNN. Additionally, it also employs Bi-directional Long Short-Term Memory (BLSTM) (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) recurrent neural networks, to model long-term temporal information. The approach is evaluated on a number of databases to compare its performance with other state-of-art-methods. A detailed experimental analysis of various system components and parameters is also provided to demonstrate the benefits and

find any possible limitations of the proposed system.

## 4.1 Methodology

The proposed approach uses small rectangular image regions and corresponding binary image masks to learn the relevant appearance and shape features, respectively. A sequence of consecutive images are used in order to model the dynamics. A transformed sequence of image regions and binary image masks are used as input to train a CNN. The dynamic features learnt from this CNN are further used for training a BLSTM neural network. The output from this BLSTM neural network serves as the final decision value for the occurrence of an AU. This section describes each component of the system in detail.

### 4.1.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a special class of feed-forward artificial neural networks which employ Convolutional layers among various other types of layers such as Pooling layers, fully-connected layers, activation function layers, etc. A Convolutional neural network usually consists of multiple layers of these types stacked one on top of another. An initial input is fed into the first layer which is processed and the output is fed into the next layer. This process is repeated for each layer and the output of the last layer is taken as the network prediction value. The most common types of layers often used in CNNs are as follows:

- **Convolutional layer:** This layer applies convolution operation defined by a set of trainable filters (weight matrices), to the input feature maps. These filters usually have a small window size (receptive field) but they fully extend along the the depth of the input feature maps (see Fig. 4.1). Convolution with each filter produces a 2 dimensional output activation map. These sets of output activation maps are used as input to the next layer.
- **Pooling layer:** This layer is used for spatial downsampling of the input feature maps. The most common type of pooling operation is the max-pooling. In max-pooling, each of the input feature maps are divided into equally sized overlapping

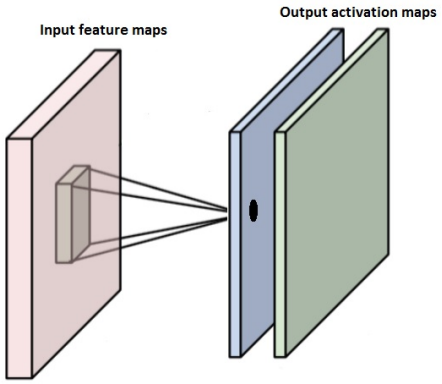


Figure 4.1: Working of a Convolutional layer in CNN.

windows and the maximum value of the feature in each window is selected to compute the downsampled map. It does not have any learnable parameters. Pooling operations reduce the number of parameters in a network and consequently help in reducing the computation time and also avoiding overfitting. It also helps in providing some degree of translation invariance to the network.

- **Fully-connected layer:** This type of layer can be viewed as a special case of Convolutional layer in which the receptive field size of each filter is equal to the size of input feature maps. In this special case, the convolution operation turns into an inner-product of the input feature volume with each filter. The output of a fully-connected layer is thus a one-dimensional vector of length equal to the number of filters used. The output of fully-connected layers do no have any spatial information and are used for computing high level features/prediction. These usually constitute the last few layers of a CNN.

- **Activation function layers:** This kind of layer is used for introducing non-linearity into the output of convolutional layers and fully-connected layers. The most common type Activation function layer is the Rectified Linear Unit (ReLU) layer. It applies the ReLU function $f(x) = max(0, x)$ to each element of the input feature map. The output activation maps have same size and depth as that of the input. Other kinds of activation functions that are used include the Sigmoid function $f(x) = (1 + e^{-x})^{-1}$ and the hyperbolic tangent function $f(x) = tanh(x)$. The use of ReLU is usually preferred over other activation functions because it helps in training the network several times faster (Krizhevsky et al., 2012b). Like the pooling layer, this layer also does not have any learnable parameters.

A loss function is used to compute the prediction error $E$ from the network outputs and the ground truth labels. The parameters of the network are learnt so as to minimize the error $E$. The learnable parameters in a CNN (filter weights and biases) are learnt using the backpropagation algorithm (Werbos, 1974) in conjunction with optimization methods such as gradient descent. In backpropagation, the error calculated at the output layer is propagated backwards until each unit in the network has an associated error. It is based on the application of chain rule and is used to calculate the gradient $\nabla E(\mathbf{w})$ where $\mathbf{w}$ is a vector of all learnable parameters of a CNN. For a detailed description of the backpropagation algorithm, readers are referred to Bishop (1995). The gradient $\nabla E(\mathbf{w})$ can be used to find the optimum $\mathbf{w}$ using gradient descent. At each iteration, the update in $\mathbf{w}$ is given by,

$$\Delta \mathbf{w} = -\alpha \nabla E(\mathbf{w}) \tag{4.1}$$

where $\alpha$ is the step-size (learning rate).

**CNN architecture of the current system:** CNN based classifiers currently provide state-of-the-art performance on a number of computer vision based tasks such as object classification (Girshick et al., 2014), face recognition (Taigman et al., 2014; Parkhi et al., 2015), etc. However, in contrast to traditional CNN architectures which usually have only one set of inputs, this approach proposes a new architecture which is designed specifically for the dual set of inputs (appearance and shape) used by this method.

The CNN architecture that we use in this work is shown in Fig. 4.2. It has 2 input streams, one for appearance representation $A$ and another for shape representation $S$. In each stream, the inputs are first passed through a separate convolutional layer ($Conv1a$ and $Conv1b$) which consists of 32 filters of spatial size $5 \times 5$. This convolutional layer is followed by a max-pooling layer of size 3x3. The outputs from both the streams are merged together after max-pooling, into a single stream. This merged stream consists of 2 more convolutional layers and 1 fully connected layer. The first convolutional layer in the merged stream ($Conv2$) consists of 64 filters of spatial size $5 \times 5$. The second convolutional layer in this stream ($Conv3$) consists of 128 filters of spatial size $4 \times 4$. The fully connected layer ($FC$) has 3072 units, which is followed by the output layer. In this network, Rectified Linear Unit (ReLU) is used as the activation functions, after each convolutional and fully-connected layer.

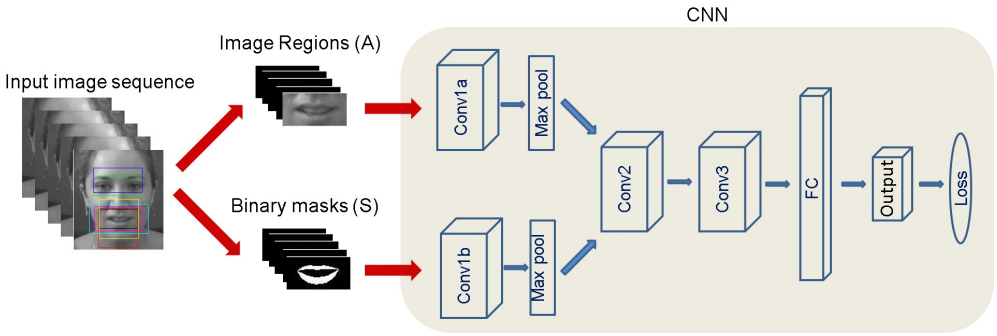


Figure 4.2: A graphical overview of the inputs to CNN and its architecture: The colored rectangles in the input image sequence show the different image regions selected for different AUs. Here, the extraction of image regions (A) and binary masks (S) for AU 12 is shown. These are used as input to the train the CNN. Loss function here is the softmax loss for occurrence detection and mean-squared error (MSE) for intensity estimation.

| Layer | filter size | Input feature-map size | No. of filters | Stride |
|---|---|---|---|---|
| Conv1 | 5x5 | 40x80 | 64 | 1 |
| Pool | 3x3 | 36x76 | - | 2 |
| Conv2 | 5x5 | 17x37 | 64 | 1 |
| Conv3 | 4x4 | 13x33 | 128 | 1 |
| FC | 10x30 | 10x30 | 3072 | - |
| Output | 1x1 | 1x1 | 1 | - |

Table 4.1: CNN architecture parameters.

### 4.1.2 Bidirectional Long Short Term Memory

Most existing CNN based methods for AU detection do not use temporal information. Those which do (e.g. Jung et al. (2015)), use only a short sequence of images in a fixed time window. This limits the access to temporal information within a specific time window only. On the other hand, different facial AUs can occur over different time scales. For e.g., Smile (AU12) can last for much longer time as compared to blinks (AU45) which occurs only for a short interval of time (See Fig. 4.3). Even within a specific AU, the variation in its duration of occurrence can be quite large.

In order to learn temporal features over longer and variable time frames, a recurrent neural network architecture known as Bidirectional Long Short-Term Memory (Graves and Schmidhuber, 2005), is used. A BLSTM network is an extension of LSTM network capable of learning information both in the forward temporal direction as well as in the backward direction. An LSTM network is basically a recurrent neural network in which the

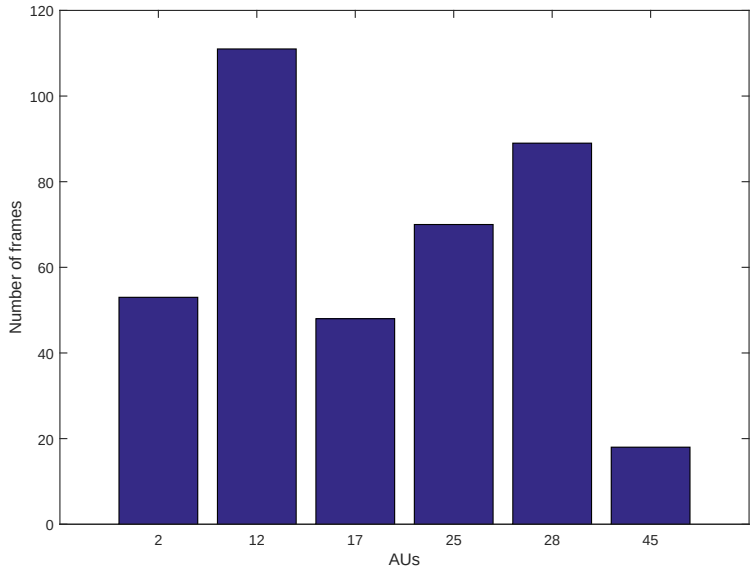


Figure 4.3: Average duration of each AU in the SEMAINE dataset.

hidden nodes are specialized memory blocks, access to which is controlled by number of multiplicative gates. Fig. 4.4 shows the different components of a LSTM memory block. Each memory block has one or more self-connected cells and 3 gates: input, output and forget gate which provide read, write and reset functions for the memory cells. Traditional RNNs suffer from the vanishing/exploding gradient problem (Hochreiter et al., 2001) due to gradient propagation across many recurrent layers each corresponding to a single time step. The multiplicative gates enable LSTM cells to avoid the vanishing gradient problem by providing the functionality to store and access information over extended periods of time. Given an input sequence $\mathbf{x} = x_1, ..., x_T$, a forward pass in an LSTM network is implemented according to the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{4.2}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4.3}$$

$$j_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{4.4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot j_t \tag{4.5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{4.6}$$

$$h_t = o_t tanh(c_t) \tag{4.7}$$

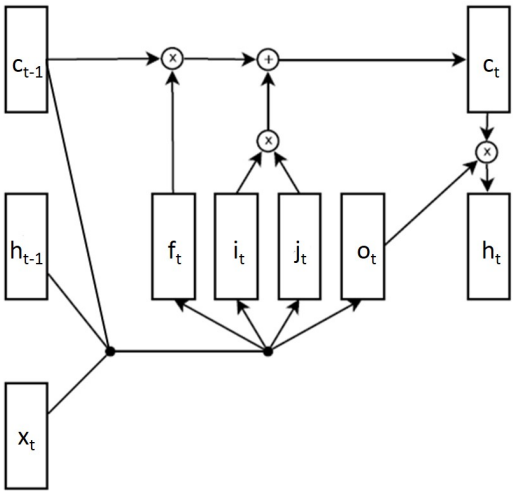


Figure 4.4: Working of a LSTM recurrent neural network. Image inspired from Jozefowicz et al. (2015). i, o and f represent the input, output and forget gate respectively. h and c represent the hidden states and cell states respectively.

$$y_t = W_{hy} h_t + b_y \tag{4.8}$$

where $h$ is the hidden vector sequence, $y$ is the output vector sequence, $\sigma$ is the logistic sigmoid function, $c$ is a vector of cell activations, $i$, $f$ and $o$ are the activations of input,forget and output gate respectively. $j$ is the activated input to the LSTM cell. $W_{xi}, W_{hi}$ and $W_{ci}$ are the weight matrices between input gate and the input sequence, hidden units and cells respectively. Similarly, $W_{xf}$,$W_{hf}$ and $W_{cf}$ are the matrices of weights between forget gate and the input sequence, hidden units and the cells respectively. Similar is the naming convention for $W_{xc}, W_{hc}, W_{xo}, W_{ho}, W_{co}$ and $W_{hy}$.

In a Bidirectional LSTM (BLSTM), there are 2 separate hidden layers. One hidden layer gets trained in the forward time direction, while the other layer gets trained in the reverse time direction. At the output layer, features from both the hidden layers are concatenated to compute the output. This enables learning information from both past and future. LSTM and BLSTM networks can be trained using the back-propagation through time (BPTT)(Werbos, 1990) algorithm.

A CNN model is trained first for each AU. From these trained CNNs, the output after the fully connected layer ($FC$), is extracted for each training/test instance. Sequences of these CNN feature vectors corresponding to all the frames in a video, are used as input to train BLSTM networks (one for each AU). The BLSTM used for this purpose had a single hidden layer consisting of 300 units. The output from this BLSTM network serves

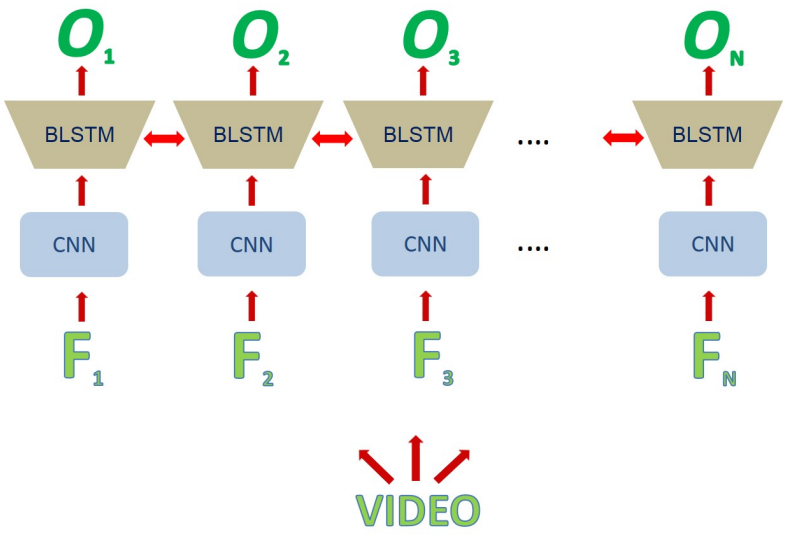


Figure 4.5: Training using a combination of CNN and BLSTM. The CNN features extracted from the FC layer are used as input to BLSTM neural network. Here the input video is first split into overlapping sequences $F_i$ each consisting of $2n+1$ consecutive frames. From each of these sequences, images regions and binary masks are extracted which are used as input to the CNN. The output decision values corresponding to $F_i$ is denoted as $O_i$ here.

as the final decision value for the occurrence/intensity of an AU.

### 4.1.3 Face tracking and landmark detection

Given a sequence of images or a video, the proposed approach starts by first detecting the face and locations of a set of unique facial landmarks, in each image/frame of the video. Face tracking and landmark detection has been a highly researched topic (Zhu and Ramanan, 2012b; Mathias et al., 2014; Xiong and De la Torre Frade, 2013) and current face tracking algorithms can track faces and their landmarks highly reliably. This system uses the ICCR face tracker (Sánchez-Lozano et al., 2016) which uses a combination of cascaded regression and incremental learning to track facial landmarks in a video. The set of 68 facial landmarks used in this approach is shown in Fig. 4.6.

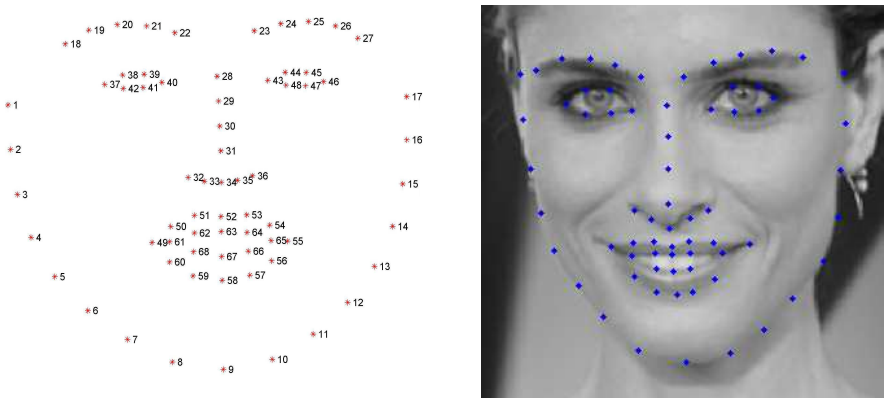

Figure 4.6: Set of 68 fiducial facial landmarks.

### 4.1.4 Face Registration

The orientation and scale of the tracked faces may vary between different frames of the videos. This can increase the intra-class variance which is undesirable. In order to minimize the intra-class variance due to multiple orientations and scale of the faces in different images, a face registration step is performed. In this step, the shape of all faces are registered against a reference face-shape.

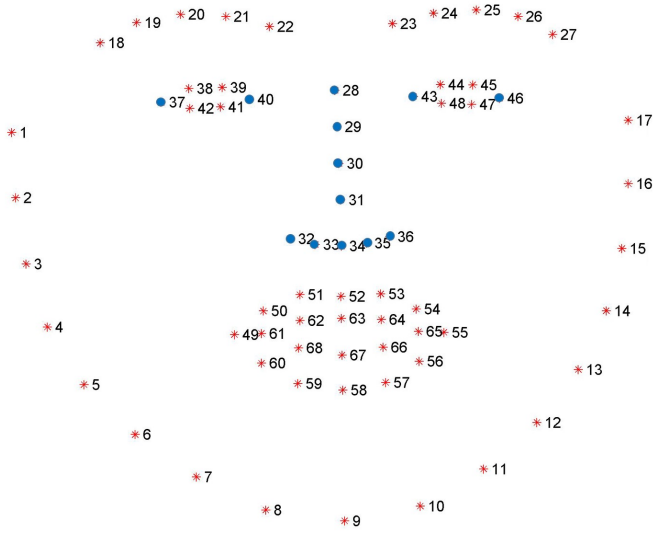

Figure 4.7: Set of stable facial landmarks (denoted with blue circles). These facial landmarks are invariant to changes in facial expressions.

A set of stable facial landmarks are selected to encode the shape. Stable facial landmarks are those landmarks whose locations are invariant to facial expressions. The set of stable facial landmarks used in this system is shown in Fig. 4.7. The shape defined by the mean locations of these stable facial landmarks in all training images is taken as the ref-

erence face shape. A shape preserving geometric transformation (Procrustes transform) is computed for other faces in the dataset. This transformation involves a combination of translation, rotation and uniform scaling so as to minimize the differences in locations between the landmarks of the current face and the reference face. If $X$ is an $n \times 2$ matrix consisting of the Cartesian coordinates of stable facial landmarks of a face, the transformed matrix is given by:

$$Y = sXT + C \tag{4.9}$$

where $s$ is a scaling parameter, $T$ is the orthogonal rotation matrix and $C$ is the matrix of translation parameters. Procrustes transform involves finding parameters $s$, $X$ and $c$ so as to minimise:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - Z_{ij})^2 \tag{4.10}$$

where $Y$ represents the transformed face shape and $Z$ represents the reference face shape. The transformation parameters obtained for each face are applied to the corresponding face image and landmarks to get the registered face image and shape (landmarks).


Figure 4.8: Face registration step using facial landmarks on the eye corners and the nose. It reduces the intra-class variance in face images.

### 4.1.5 Image Regions

Learning facial action unit models can be a difficult task because only a small part of the face is responsible for the occurrence of a specific Action Unit. Noise from other (non relevant)parts of the face can inhibit the learning of models with low generalization error. Implicit learning of facial regions which are responsible for a particular Action Unit is as difficult task for current machine learning algorithms due to the high dimensionality of the input image data and the limited amount of training examples available. However, from domain knowledge one can easily infer which regions show maximum change in

appearance and/or shape, for each action unit. For e.g. action units 12 (lip corner puller), 25 (lips part) and 26 (jaw drop) are all confined to the mouth region. Similarly, action units 1 (inner brow raiser), 2 (outer brow raiser) and 4 (brow lowerer) are all confined only to the eye region. Hence, to block the noise coming from irrelevant parts of the face and to guide the CNN in learning features which generalize well, the relevant image image region for particular Action Unit is pre-selected according to domain knowledge.

In order to define an image region, a small set of facial landmarks is selected according to domain knowledge. The location of these selected facial landmarks correspond to the region of the face where the target AU produces maximum change in shape and appearance. The mean location (centroid) of the selected facial landmarks is taken as the centre around which a rectangular image region of a fixed width $w$ and height $h$, is defined. The values of $w$ and $h$, which can be different for each AU, are selected so as to construct a minimum sized region which covers most of the appearance and shape variation caused by the activation of an AU. These were chosen through visual inspection of average image computed from from all the faces where a specific AU is active. The rectangular image region is cropped from the original image and is denoted as $f$ (see Fig. 4.9). A list of AUs and the corresponding facial landmarks selected for defining the image regions, are shown in Table 4.2. The rectangular regions selected for each AU are also shown in Fig.4.10.

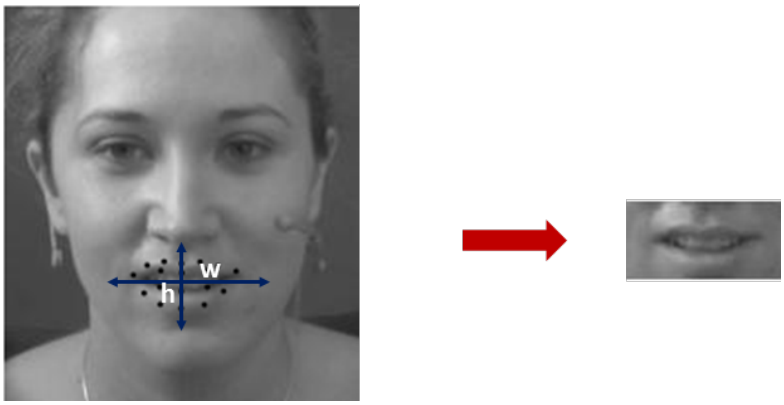


Figure 4.9: Construction of image region for AU 25.

Most existing CNN based approaches use whole face images for learning and predicting facial AUs. This is in contrast to the proposed methodology which uses local image regions. Using the entire face image will require larger amount of training data as the input dimensionality will be very high. Using whole face images when only limited amount of training data is available, may cause overfitting. However, one of the drawbacks of

| Action Unit | Facial landmarks | w (in pixels) | h (in pixels) |
|---|---|---|---|
| 1, 2, 4 | 18-27,46-48,43,40-42,37 | 100 | 50 |
| 6, 12, 14, 20, 26 | 49-55,65,64,63,62,61; 55-60,49,61,68,67,66,65 | 100 | 50 |
| 15, 23, 25, 28 | 49-55,65,64,63,62,61; 55-60,49,61,68,67,66,65 | 80 | 40 |
| 5, 7, 45 | 37-42; 43-48 | 100 | 40 |
| 17 | 7-11,55,53,51,49 | 80 | 80 |
| 9 | 28-31 | 60 | 80 |
| 10 | 49-55,65,64,63,62,61; 55-60,49,61,68,67,66,65 | 80 | 80 |

Table 4.2: Facial landmarks, width $w$ and height $h$ used for defining image regions of each AU. The width $w$ and height $h$ are given for a registered face image of size $180 \times 200$ pixels.

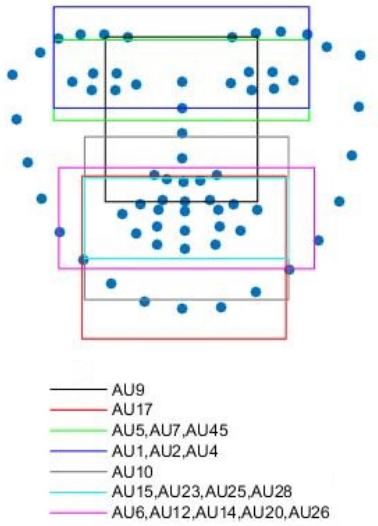


Figure 4.10: Visualization of the rectangular facial regions selected for different AUs.

this method is that it is not possible to learn the correlations between AUs which occur in different image regions (as defined above) of the face. Learning such correlations can be helpful because it is well known that certain AUs frequently co-occur together while certain other AUs rarely occur together. An alternative to this would be to use the entire face image as input to the CNN, which is usually the norm in most AU recognition systems. However, it will be shown in the experimental section that using the proposed image regions provides a significant overall gain over using the entire face image. Another limitation of this approach is that the image regions are defined manually for each AU, which may not give the most optimum results. There has been some recent work which attempts to learn such local regions from the data itself. For e.g. Sariyanidi et al. (2017) proposed an unsupervised technique to learn localised bases whose linear combination can be used to represent facial expressions. However, more research is required

to develop such techniques especially for supervised deep learning approaches.

### 4.1.6 Binary masks

The activation of a facial AU causes changes in the appearance and shape of facial parts. Change in appearance can be dominating for some AUs (e.g. AU1, AU2) while for some other AUs change in shape can be highly discriminative (e.g. AU25, AU26). Hence, learning shape features is equally essential for any method which aims to learn accurate models for all AUs. However, learning shape variations implicitly while directly training for AUs is difficult because it not only involves learning which parts of the face are responsible for shape variation during the activation of a specific AU, but it also requires precise localization of those parts in each face image. Facial landmark detectors on the other hand are trained specifically for the task of localising individual parts of a face and current state-of-the-art detectors can do this task at a high precision (at least for near frontal faces). The parts of the face whose shape can be discriminative for the activation of an AU can be specified from domain knowledge while leaving the task of localising those parts to facial landmark detectors. The traditional way of encoding shape is to compute hand-crafted geometric features from the locations of facial landmarks (see section 3.4.1). However, as observed in the field of object classification/detection and segmentation, learning features using deep CNNs can vastly improve performance. Therefore, this work employs CNNs to learn the shape features instead of using hand-crafted features.

In order to guide the CNN to encode shape information of a relevant part of the face, a binary mask $b_i$ is computed corresponding to an image region $f_i$. To compute the binary masks, the facial landmarks selected for defining the image regions (see Table 4.2) are joined together to form a set of polygons. The order in which the landmarks are joined is defined in such a manner so that the resulting polygons roughly represent the segmentation of one or more distinct parts of a face. For e.g., in case of AU12, the landmarks on the lips are joined so that the resulting polygons give a rough segmentation of the upper and lower lips. For each AU, the order in which the landmarks are joined is given in Table 4.2. The binary mask image $b_i$ corresponding to the image region $f_i$ is computed by setting all the pixel values of $f_i$ which lie inside a polygon to $1$ and all the pixel values which lie outside a polygon to $0$ (see Fig. 4.11). This rasterized

representation of shape computed from facial landmarks enables it to be used directly in a CNN without any change in the architecture. It not only makes full use of the accurate landmark localization provided by the detectors but it also allows CNNs to learn the optimum shape features instead of using hand-crafted geometric features. This fact has been tested in the experimental section where a CNN is learnt using binary masks only and its performance was compared to a hand-crafted geometric feature based approach (Valstar et al., 2015c).

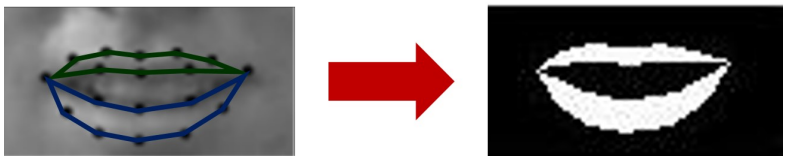

Figure 4.11: Construction of binary mask for AU 25.

### 4.1.7 Dynamic encoding

Facial Action Units can be defined as the movement (actions) of individual or group of facial muscles. By virtue of their definition, facial AUs have a strong temporal aspect within them and can provide discriminative features for recognition of any facial action unit (Valstar and Pantic, 2012a; Almaev and Valstar, 2013). Short term temporal dynamics (typically involving only few frames) can not only help in localising which facial region is temporally active but the modes of temporal variation can be used to discriminate one AU from another.

To encode short term temporal information, image regions $\{f_{t-n}, ..., f_{t+n}\}$ and their corresponding binary masks images $\{b_{t-n}, ..., b_{t+n}\}$ are extracted from a sequence of $2n+1$ consecutive frames of a video centred at the current frame $t$. The resulting sequence of image regions are transformed to a sequence $A = \{A_i\}$ where

$$A_i = \begin{cases} f_i, & \text{if } i = t \\ f_i - f_t, & \text{otherwise} \end{cases} \tag{4.11}$$

Similarly, the sequence of binary mask images are transformed to a sequence $S = \{S_i\}$

where

$$S_i = \begin{cases} b_i, & \text{if } i = t \\ b_i - b_t, & \text{otherwise} \end{cases} \tag{4.12}$$

The resulting image sequences $A$ and $S$ are used as input to a deep Convolutional Neural Network (CNN). This is in contrast to the previous approaches (Liu et al., 2015; Jung et al., 2015) which directly use the images within a time window around the current frame. This method of transforming the image sequences by taking difference from the current frame, makes it easier to learn the dynamics in a CNN framework. One limitation of this approach is that it requires accurate tracking of facial landmarks. The current state-of-the-art trackers can provide very accurate and stable stacking at least for frontal and near frontal faces. However, in case the landmark tracking fails, the proposed method is likely to give incorrect results.

It should be noted that here short term dynamics refers to temporal information contained within a time window of up to one second (typically 0.1 to 1 sec). Anything longer than that will be referred to as long term dynamics. Short term dynamics can therefore capture parts of an AU onset, peak or offset. However, in order to capture the entire sequence of onset-peak-offset of an AU, long term dynamics will be required which is learnt using BLSTM.

### 4.1.8 Training with CNN and BLSTM

The sequence of image regions $A$ and the sequence of binary shape masks $S$ for each training example are fed as input to the CNN described in section 4.1.1. The $2n + 1$ images in each of the sequences $A$ and $S$ are stacked together in two separate groups which are fed into two input streams of the CNN as $2n+1$ channel inputs each. To avoid overfitting, a regularization technique known as dropout (Srivastava et al., 2014), is used in which the output of a neuron is randomly set to zero with a certain probability $p$. This system used dropout on the fully connected layer ($FC$) with probability $p = 0.2$.

In order to train the network, two kinds of loss functions are used:

1. For AU occurrence detection, which involves classification of each frame into

whether an AU is active or inactive, a softmax loss function is used. It is given by,

$$L_s = -\sum_i \sum_j y_{ij} log \left( \frac{e^{z_{ij}}}{\sum_k e^{z_{ik}}} \right) \tag{4.13}$$

where $y_{ij}$ denotes the ground truth labels and $z_{ij}, z_{ik}$ denotes the output of the network. The index $j$ runs over all the possible ground truth classes, index $k$ runs over all the outputs of the network and index $i$ runs over all the training examples.

2. For AU intensity estimation (regression problem), mean squared error (MSE) is used. It is given by,

$$L_m = \frac{1}{N} \sum_i^N \sum_j (y_{ij} - z_{ij})^2 \tag{4.14}$$

where $y_{ij}$ and $z_{ij}$ represent the ground truth labels and output of the network respectively. Index $j$ runs over all the possible ground truth classes and index $i$ runs over all the training examples.

The training data is normalized so as to have a zero mean and one standard deviation for each pixel across all training examples. The networks are trained using mini-batch gradient descent with momentum. A batch size of 100 training examples are used while the momentum and learning-rate parameters were fixed to 0.9 and 0.01 respectively. The weight updates are calculated using the back-propagation algorithm. This system uses the MatConvNet library (Vedaldi and Lenc, 2015) for training the CNN models.

The trained CNNs are used as feature extractors for learning BLSTMs. The output of the trained CNN after the $FC$ layer is extracted for each training example. This results in a 3072 dimensional dynamic CNN features for each example. Sequences of these feature vectors corresponding to all frames in a video, are used as input to a BLSTM network. For each AU, a separate BLSTM network is trained. Each BLSTM network consists of one hidden layer of size 300 units and one output layer of size 1 unit. As in the case of CNNs, the softmax loss function is used for AU occurrence detection models and MSE loss function is used for AU intensity models. This system used the Munich Open-source CUDA Recurrent Neural Network Toolkit (CURRENNT) (Weninger et al., 2015), for training the BLSTM networks. The training was done using batch gradient descent with momentum and learning-rate parameters fixed to 0.9 and 0.00001 respectively.

## 4.2 Evaluation

This section describes the experiments conducted to evaluate different components of the proposed system and to compare its performance with other state-of-the-art approaches. An overview of the datasets and the performance metrics used in the experiments is given first followed by the description of various experiments and their corresponding results.

### 4.2.1 Datasets

The proposed approach was evaluated on the following four different databases:

- **SEMAINE:** The SEMAINE database (Mckeown et al., 2012) is a part of the FERA-2015 Challenge dataset (Valstar et al., 2015a). It was recorded to study social signals that occur when people interact with virtual humans. It consists of videos in which users are interacting with emotionally stereotyped characters. A total of 6 Facial Action Units are labeled for each frame in the videos. The dataset is divided into a fixed training, development and test set. The partitioning is subject independent , i.e. the subjects present in the training set are not present in the test set and vice versa. The training partition consists of 16 sessions, the development partition has 15 sessions, and the test partition has 12 sessions. There are approximately 48,000 images in the training partition, 45,000 in the development and 37,695 in the test partition.

- **BP4D:** The BP4D database (Zhang et al., 2014b) is also a part of the FERA-2015 Challenge dataset. It consists of recorded videos in which the subjects are responding to emotion elicitation tasks. Like SEMAINE, BP4D is also divided into a fixed set of training, development and test data. The training set consists of 21 subjects while the development and the test set consists of 20 subjects each. There are 8 sessions for each subject. In total, the training partition contains 75,586 images, the development contains 71,261 images and the test contains 75,726 images. Each of these images are annotated with 11 Action Units. For 6 of these Action Units, only occurrence labels are available. For the other 5 Action Units occurrence as well as intensity levels are available.

- **DISFA:** Denver Intensity of Spontaneous Facial Action (DISFA) database (Mava-

dati et al., 2013) is a publicly available database which consists of non-posed facial expressions from 27 subjects. The videos in this dataset consist of facial images of subjects recorded while watching an emotive video stimulus. Each video frame in this dataset is manually annotated with occurrence and intensities of 12 facial AUs. The intensities are annotated on a scale of 0-5, 0 denoting the absence of a AU and 5 denoting the highest intensity for an AU. In total, there are approximately 130,000 video frames in this dataset.

- **CK+:** The Cohn-Kanade (CK+) database (Lucey et al., 2010) is another publicly available database which is often used for evaluating emotions recognition algorithms in predicting 6 prototypic basic emotions. This database consists of 593 image sequences from 123 subjects. The sequences consists of both posed and spontaneous expressions displayed by the subjects. Each sequence starts with a neutral face and it progressively increases in the expression intensity. The last frame which corresponds to the peak expression intensity, is annotated with one of 7 basic emotions namely, Anger, Disgust, Contempt, Happiness, Sadness ans Surprise.

### 4.2.2 Performance metrics

A number of performance metrics are used for evaluating the performance of the proposed system. 2AFC and F1 measures are used for evaluating occurrence detection models while the ICC measure is used for evaluating intensity estimation models. Each of these performance measures are described in detail below:

- **2AFC:** In all the experiments, wherever possible we used the 2 Alternative Forced Choice (2AFC) score as a measure for evaluating the performance for the task of AU occurrence detection. The 2AFC score is a good approximation of the area under the receiver operator characteristic curve (AUC) (Tyler and Chen, 2000) and is considered to be a good measure for both balanced and imbalanced datasets. It is defined as follows:

$$2AFC(\hat{Y}) = \sum_{i=0}^{n}\sum_{j=0}^{p}\sigma(P_j, N_i)\frac{1}{n \times p}, \tag{4.15}$$

$$\sigma(X,Y) = \begin{cases} 1, & \text{if } X > Y \\ 0.5, & \text{if } X == Y \\ 0, & \text{if } X < Y \end{cases}$$

where $\hat{Y}$ is a vector of output decision values from a classifier, $P$ and $N$ are subsets of $\hat{Y}$ corresponding to all positive and negative instances, respectively. $n$ is the total number of true negatives and $p$ is the total number of true positives.

- **F1:** Although 2AFC measure is used for most of the experimental evaluations, for some databases F1 measures are also computed for a fair comparison with other methods reporting on the same dataset. The F1 measure can be defined as follows:

$$F1 = \frac{2.Precision.Recall}{Precision + Recall} \tag{4.16}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.17}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.18}$$

where TP is number of true-positives, FP is the number of false-positives and FN is the number of false-negatives returned by a classifier.

- **ICC:** For intensity estimation of facial AUs, the Intraclass Correlation Coefficient (ICC) measure (Shrout and Fleiss, 1979), is used for evaluating the performance. Given the two sources of coding $y_{i1}$ and $y_{i2}$, the ICC between these two sources is defined as,

$$ICC = \frac{W - S}{W + S} \tag{4.19}$$

where $W$ is the within-target mean squares and $S$ is residual sum of squares. These can be computed as follows,

$$W = \sum_{i=1}^{n} \sum_{j=1}^{2} \frac{y_{ij} - \overline{y}_i}{n} \tag{4.20}$$

$$S = \sum_{i=1}^{n} (y_{i1} - y_{i2})^2 \tag{4.21}$$

where $\overline{y}_i = \sum_{j=1}^{2} y_{ij}/2$. In the present case, the two sources $y_{i1}$ and $y_{i2}$ are the ground truth intensity labels and the predicted intensities from the learnt model.

### 4.2.3 Experiments

A number of experiments were conducted to demonstrate the effectiveness of the proposed model and compare its performance with other state of the art methods. The first set of experiments evaluates the contribution of different components of this system and the sensitivity of various parameters on the performance. In the second set of experiments, the performance of this approach is compared with other state-of-the-art approaches, on the task of AU occurrence detection and intensity estimation. It should be noted that in all of the following experiments no data augmentation techniques (e.g. random cropping, image rotation, etc.) were used to train the models. This was because all the training examples were pre-processed by doing a face registration step which removes most of the effects (translation, rotation, scaling) added by the data augmentation techniques.

**Effect of using sequence of image regions and binary shape masks:**
In order to demonstrate the effectiveness of using image regions and binary shape masks, a number of experiments were conducted on the SEMAINE dataset. To demonstrate the advantage of using image regions as opposed to full image of faces, two separate baseline models were computed. The first model consists of a CNN similar to the CNN in Fig. 4.2, except that the input to this CNN is the full image of a face (aligned) defined by the face bounding box and the temporal window parameter $n = 2$. There are no image regions (defined by facial landmarks) or binary masks as input to this CNN. This baseline model is denoted as $\text{CF}_{n=2}$. The second baseline model ($\text{CR}_{n=2}$) uses the same CNN architecture but uses image regions (see section 4.1.5) as input. It should be noted that BLSTM was not used for this set of experiments and the performances are evaluated based on CNN output only. Fig.4.12 shows a comparison of the performance of these two models. In this plot we can observe that the performance of $\text{CR}_{n=2}$ is higher than that of $\text{CF}_{n=2}$, for AU17, AU25 and AU45. For other AUs, the performance didn't change significantly. On average, the performance was significantly higher for $\text{CR}_{n=2}$ indicating that in most cases, using local regions of faces is preferable against full face images. However, in some cases where an AU is strongly correlated to some other AU which occurs in a completely different region of the face, using only a local image region (as input) may not be the best approach. In such a case the network will not be able to learn these correlations which could have potentially reduced the chances of detection errors. This might be the case with AU28 where the performance is slightly higher when using

full face images.

To demonstrate the contribution of shape encoded by binary masks and its joint modelling with appearance represented by local image regions, the performance of three kinds of baseline models were compared on the SEMAINE dataset. The first baseline model is $CR_{n=2}$ computed earlier which uses only appearance represented by image regions. The second baseline $CM_{n=2}$ uses shape information only encoded by binary masks. The baseline model $CRM_{n=2}$ learns both appearance and shape using the two stream CNN described in Section 4.1.1. Fig.4.13 shows the performance comparison of these baseline models. In this figure, it can be observed that the effect of using shape and appearance information varies significantly with AUs. For instance, in case of AU2, appearance information dominates its performance and addition of shape information does not result in any significant change. This was expected considering the fact that when AU2 is activated, the shape of eye brow changes little compared to the appearance changes caused by the wrinkles produced on the forehead. On the other hand, the exact opposite phenomenon is observed in case of AU45 where shape information dominates and addition of appearance information results in no significant change in performance. This observation can also be explained considering that for AU45 (blinks), the shape of the eye (as tracked by the landmarks on the eyes) undergoes a drastic change when AU45 becomes active (the corresponding shape in the computed binary masks change from a 6 sided polygon with finite area to almost a straight line). Similar observation can be made for AU25. For all other AUs (AU12, AU17 and AU28), it can be observed that both shape and appearance information contribute significantly to the performance. Overall, the best average performance (across all AUs) is achieved when both appearance and shape information is jointly learnt in the CNN. The average performance of the shape only model ($CM_{n=2}$) can also be directly compared with the hand-crafted geometric feature based approach by Valstar et al. (2015c) which attained an average 2AFC and F1 score of 0.19 and 0.39 respectively, on this dataset. In contrast, our model $CM_{n=2}$ achieves an average 2AFC and F1 scores of 0.80 and 0.43 respectively. This shows that the proposed method of using binary masks enables the CNN to learn more optimal shape features resulting in significantly better performance as compared to using hand-crafted shape features.

**Effect of using dynamic CNN and BLSTM:**

We also computed a number of baseline models for demonstrating the effect of using dynamics in CNN and the effect of adding BLSTM to our training pipeline. We computed

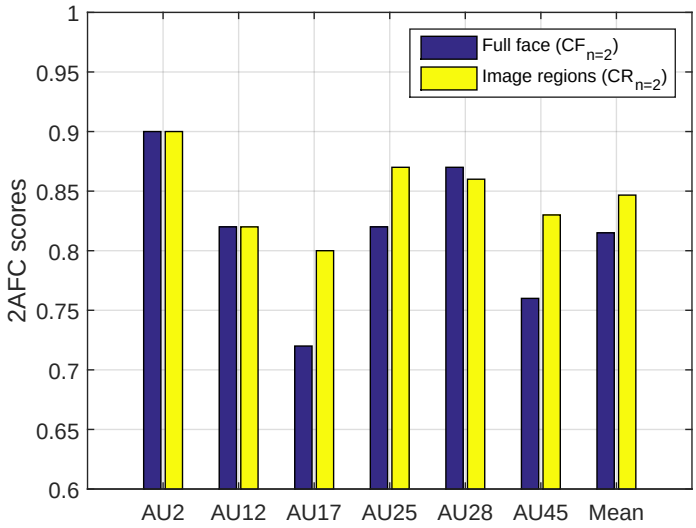


Figure 4.12: Performance comparison of full face model and local image region based model on the SEMAINE test set.

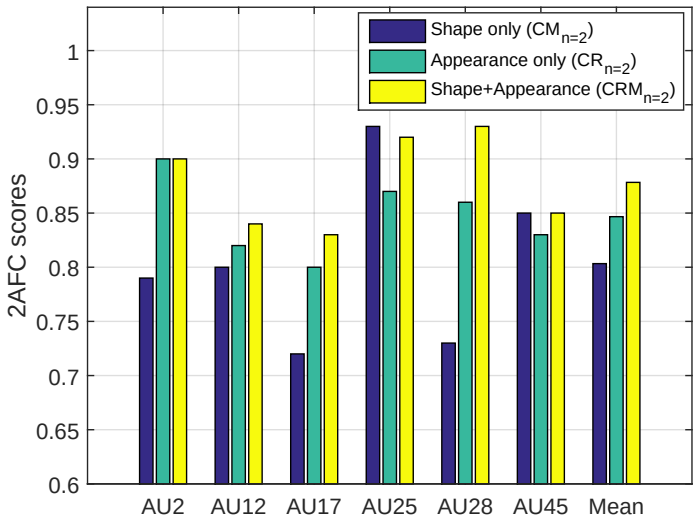


Figure 4.13: Performance comparison of shape (represented by binary masks) only models, appearance (represented by local image regions) only models and shape+appearance models on the SEMAINE test set.

the performance of our approach with BLSTM (CRML$_{n=2}$) along with 3 other methods. Our first baseline is the CRM$_{n=0}$, which does not use any temporal window in CNN and also does not use BLSTM. Hence this baseline does not use any dynamics. Our second baseline is CRML$_{n=0}$ which does not use any temporal window during CNN training but employs BLSTM after CNN training. Our third baseline is CRM$_{n=2}$ which uses a $2n + 1 = 5$ frame temporal window during CNN training, but does not use BLSTM.

Fig. 4.14 shows the relative performance of our final approach CRML$_{n=2}$ and the other 3 baseline methods. We observe that on average the performance of CRML$_{n=0}$ is higher than that of CRM$_{n=0}$, indicating that BLSTM is able to learn the dynamics resulting in an improved performance even without using a temporal window during CNN training. The performance of CRM$_{n=2}$ over that of CRM$_{n=0}$ indicates the extent to which, using a temporal window of 5 frames while training the CNN, can improve the performance even without using BLSTMs. However, the best average performance is achieved with CRML$_{n=2}$, which uses a 5 frame temporal window during CNN training and also employs BLSTM for training with features extracted from CNN. This indicates that on average, BLSTMs performs better with dynamic features (extracted using a fixed temporal window in CNN) compared to non-dynamic features (extracted using a single image in CNN). Taking a closer look at the performance of each AU model, it can be observed that for AU2, AU12 and AU17 BLSTM alone is sufficient to model the dynamics and the addition of dynamic CNN features does not change the performance significantly. On the other hand for AU25, AU28 and AU45 both BLSTM and short term dynamics learnt using CNNs, contribute to the increase in performance. A possible reason for this observation could be that in case of AU2, AU12 and AU17 the peak region (apex) lasts for a longer time and hence long range temporal information is much more beneficial for these AUs. In case of AU25, AU28 and AU45 the peak region lasts for a comparatively shorter interval of time and hence the region where appearance/shape changes occur (onset/offset) can be captured from short range temporal information.

**Effect of using image differences for dynamic encoding in CNN:**

The input sequences of image regions and shape masks were transformed by subtracting them with the current (middle) frame (as described in Section 4.1.7). In order to test the advantage of using such a transformation, two different sets of models were evaluated on the SEMAINE dataset. The first set of models uses no such transformation and the sequence of image regions and masks were directly used as input to the CNN. The

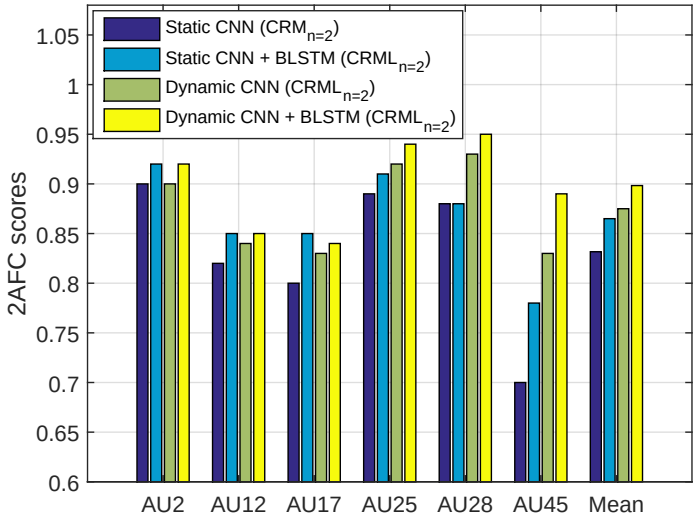


Figure 4.14: Performance (2AFC scores) comparison on the SEMAINE test set when using BLSTMs with CNNs.

second set of models use transformed sequences of image regions and masks as input to the CNN. Fig.4.15 shows a comparison of the performance of these models. In this Fig., it can be observed that when the transformation is used, the performance for AU12, AU17 and AU45 increases by 2.4%, 5.1% and 4.9% respectively. For AU2 and AU28 the performance dropped slightly by 1.1% and 0.7% respectively. Overall, the mean performance over all AUs, is 1.6% higher for the models which use transformed sequences, as compared to the non-transformed ones. In order to test the statistical significance of these results, a t-Test was also performed on the classifier outputs from the two kinds of models for each AU. It was observed that the null hypothesis (data samples are coming from normal distributions with equal mean and equal variance) was rejected for AU2, AU12, AU17, AU25 and AU45, indicating that the results are significant for these AUs (at 5% significance level). It should be noted that BLSTMs were not used for this set of experiments.

**Architectural parameters:**

In another set of experiments, we experimented with the architectural parameters of the CNN. More specifically, we evaluated the effect of adding more max-pooling (mp) layers (adding one after each convolutional layer), increasing the dropout probability (dp) at the fully connected (FC) layer and decreasing the size of FC the layer (fs). We took our proposed model CRM$_{n=2}$ which was originally trained with the parameters: {mp=1,

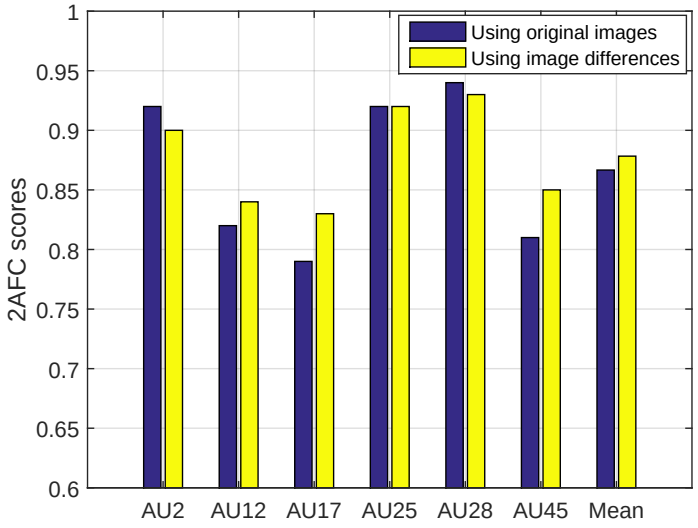


Figure 4.15: Performance comparison of AU models with and without the transformation of input sequences in CNN. The transformation is done by subtracting the frames in the sequence with current frame (described in Section 4.1.7). The performance is evaluated on the SEMAINE test set.

dp=0.2, fs=3072} and compared its performance with 4 other models each of which was trained by changing one of the parameters (mp,dp,fs). The first model was trained with parameters: mp=2, dp=0.2, fs=3072 by adding an additional max-pooling layer after the Conv2 layer. Similarly the second model was trained by adding additional max-pooling layers after Conv2 and Conv3 with parameters: {mp=3, dp=0.2, fs=3072}. In the third model, only the dropout probability was increased to 0.5, parameters being: {mp=1, dp=0.5, fs=3072}. In the fourth model, the size of FC layer was reduced to 1024, parameters being: {mp=1, dp=0.2, fs=1024}. Fig.4.16 shows the performance comparison of all the five models on the SEMAINE test set. In this plot, it can be observed that the average performance does not significantly change for each of these models except for the model where two additional max-pooling layers were used (after Conv2 and Conv3), in which case the mean performance drops by 2.7%. A possible reason for this drop could be the significant loss of spatial information that would have occurred after using 2 additional max-pooling layers.

**Sensitivity towards test frame-rate:**

In the proposed method, the CNNs are trained using image regions and binary masks (as input) extracted from a short sequence of video frames (face images). These frames have been recorded at a specific frame-rate and hence the temporal information learnt

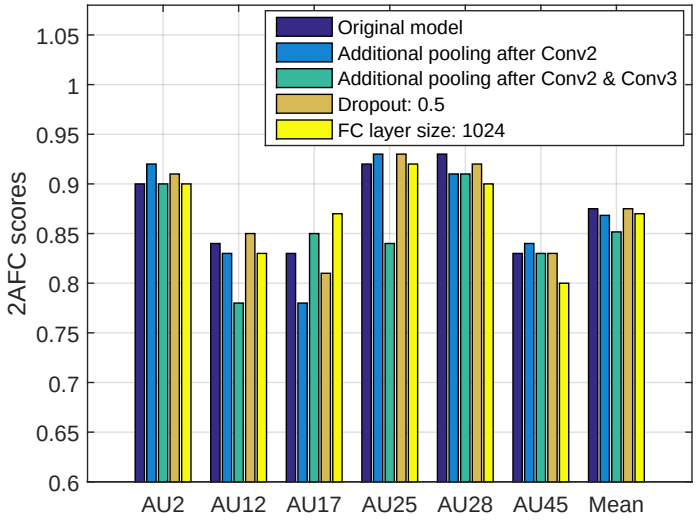


Figure 4.16: Performance comparison of models each with a different set of CNN architectural parameters. The performance is measured on the SEMAINE test set.

from these sequences could be dependent on the frame-rate of the videos used in the training stage. Therefore, if the frame-rate of the videos used during the test stage is different from the one used during the training stage, it may have a negative impact on the performance of the network. It is important to analyse the extent of this problem in order to determine how well the models will generalize if they are tested on videos recorded at a different frame-rate than the videos used for training. There have been some previous work which have also looked at this issue (for e.g. Sariyanidi et al. (2017)), as this can be an important factor if the aim is to use the trained models to detect AUs in videos where the frame-rate can be different.

To analyse this, we conducted an experiment where we tested the sensitivity of the trained models towards differences in the video frame rate during the training and test phase. In order do it, we used the AU models trained on the SEMAINE database in which the videos have been originally recorded at 50fps. The frame-rate in the test partition of the SEMAINE database was however sequentially lowered at each step by a factor of 0.5 (from the original 50fps) and each time the performance was measured in terms of 2AFC scores. A plot of the performance at each frame-rate is shown in Fig. 4.17. In this plot, it can be observed that the performance for all AUs doesn't change significantly when the frame-rate is half or even quarter of the original frame-rate (except AU45). For AU2 and AU12, the performance degrades only very slightly even for large down-sampling

of the original test videos. However for other AUs, reducing the frame-rate beyond a factor of 0.25, degrades the performance significantly, especially for AU45 where the dip in the performance is the largest. These results indicate that the performance of the AU models is robust to changes in the test frame-rate at least up to a factor of 0.5. This is important because the aim is to apply the learnt AU models to do facial behaviour analysis for detection of ADHD and ASD (see Chapter 6), where the frame-rate of the recorded videos can be slightly different.

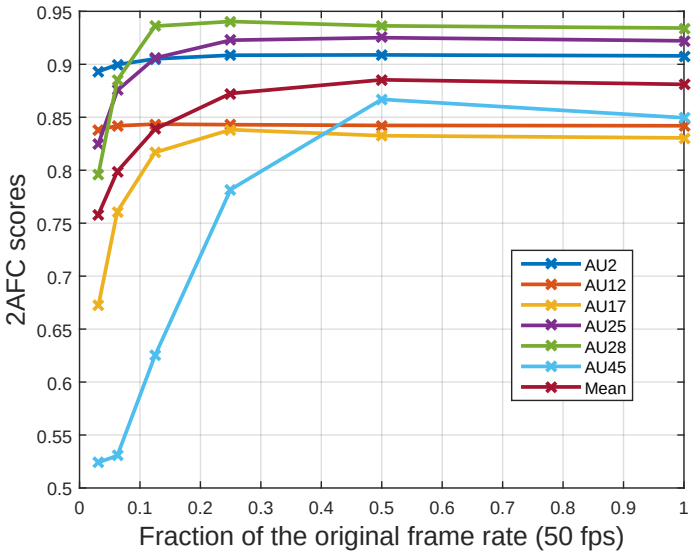


Figure 4.17: Performance of the AU models (trained on the SEMAINE database at its original frame-rate) at different frame-rates of the videos in the test partition of the SEMAINE database.

**Training separate AU models vs training together:**

The training of AU models is possible in two different ways: using a separate network for each AU or training all AUs together which belong to the same facial region. In all the previous experiments, we used separate networks to train each AU model. In order to compare the performance of the models trained in these two separate configurations, an experiment was conducted in which a network (Eye-Net) was trained to predict AUs belonging to the eye region (AU2, AU45). A separate network (Mouth-Net) was trained to predict AUs belonging to the mouth region (AU12, AU17, AU25, AU28). Due to the nature of our approach in which only a specific region of the face is used as input, it was not possible to train all AUs in a single network. A single network to learn all AUs can be used but such a network would require the entire face bounding box as input to the network because the image regions defined for one set of AUs (e.g. 12,25)

may not work for other set of AUs(e.g. 2,45) as they occur in completely different parts of the face. To minimise the number of networks to be trained but at the same time using the concept of localised image regions proposed in this work, two separate networks were trained. The input region for the first network is defined by the facial landmarks on the eyes, while that of the second network was defined by the landmarks on the mouth. These two networks for used training and predicting the AU sets [AU2,AU45] and [AU12,AU17,AU25,AU28] respectively (see Fig.4.18). These networks had 2 and 4 output units respectively, in contrast to the single AU networks which had only one output units.

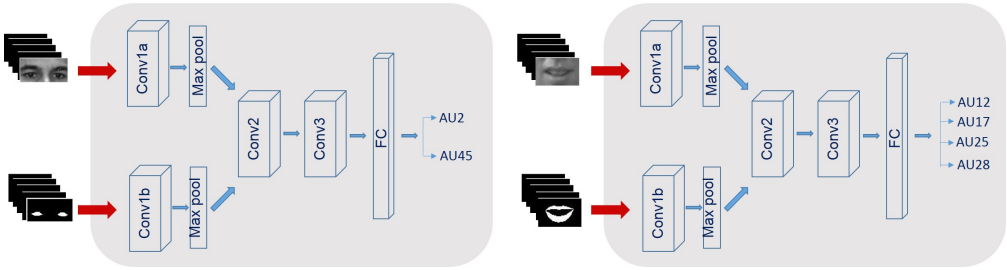


Figure 4.18: Architecture of the Eye-Net (left) and Mouth-Net (right). Eye-net takes in as input a region around the eyes to predict AU2 and AU45, while Mouth-net takes in as input a region around the mouth to predict AU12, AU17, AU25 and AU28 in the SEMAINE dataset.

Fig. 4.19 shows a comparison of the performance when the AUs belonging to the same region are trained together with the performance of the single AU networks. The performance for each of the networks was measured in terms of 2AFC scores on the SEMAINE test set. In the figure, it can be observed that for AU2, AU12 and AU25 there is a marginal increase in the performance when the AUs are trained together. However, for AU17, AU28 and AU45, there is a significant drop in the performance when training the AUs together. Overall, the average performance over all the 6 AUs in the SEMAINE database, shows a relative performance drop of approximately 3% when AUs are trained together as compared to training them using separate networks. Although having fewer networks is desirable because of low computational cost and efficient memory usage, the above results tells us that training multiple AUs together in a single network is a harder task and can result in significant performance drop for some AUs. A possible reason behind this result could be the highly varied nature of different AUs in terms of shape, appearance and especially the dynamics. Using more data could potentially overcome this problem. However, in order to do a fair comparison with other existing approaches,

using additional databases was not considered for the current set of experiments.

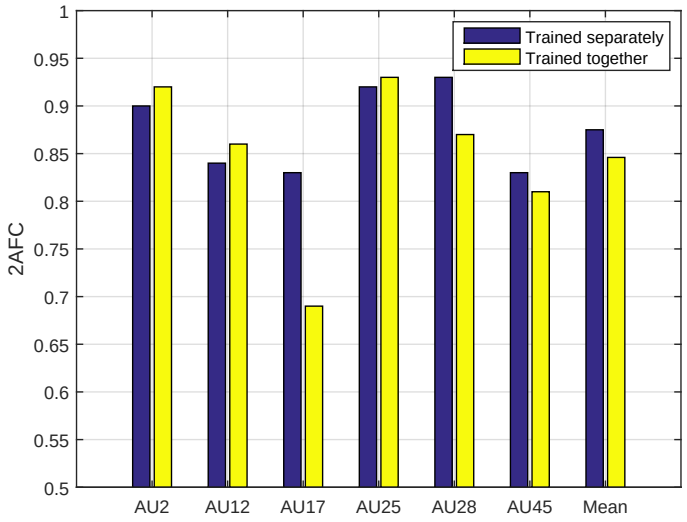


Figure 4.19: Performance comparison between models trained to predict multiple AUs using a single network, against the AU models trained using separate networks.

**Effect of time-window size**

In order to find the effect of the size of CNN time-window, a set of experiments was conducted in which the length of the sequence of image regions and binary shape masks was varied and performance was measured for each value of sequence length. For this set of experiments, the models were trained and tested on the SEMAINE database using the CNN models only (CRM). Fig. 4.20 shows the performance of the models for all the 6 AUs in the SEMAINE database (frame rate: 50 fps) in terms of 2AFC scores. The performance was tested for dynamic encoding parameter $n = \{1, 2, 5, 7, 10\}$ resulting in sequence lengths of 1, 5, 11, 15 and 21. In the Fig., it can be observed that for most AUs, the performance increases significantly as the sequence length was increased from 1 to 5. However, increasing the sequence length beyond 5, only a marginal change in the performance was observed with the average performance peaking at a sequence length of 11. It can also be observed that for different AUs, the peak performance can be obtained at different length of sequences. This was expected because the duration of activity can vary a lot from one AU to another (as seen in Fig.4.3). It is also interesting to observe that AU45 reaches peak performance at a sequence length of 15 frames, which is quite close to it's average duration (18 frames) found in the SEMAINE database. However, since the average performance increases only marginally when the sequence length is

increased from 5 to 11 and it is more efficient to have a smaller sequence length (in terms of memory and computation), we used sequence length of 5 to report performance in all other experiments.

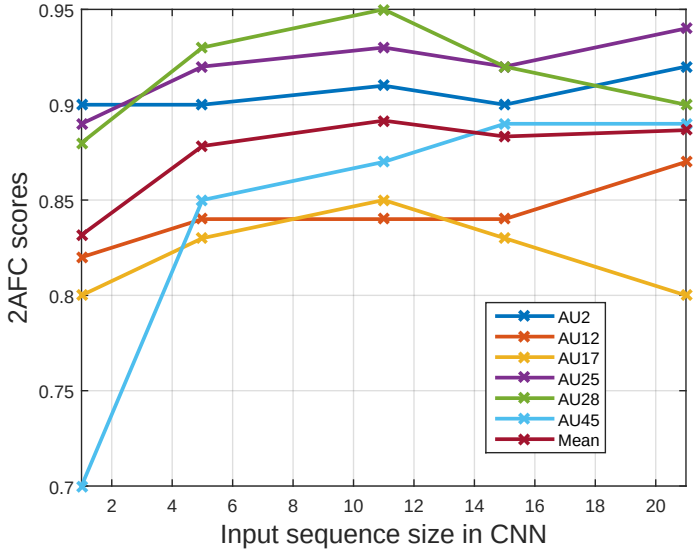


Figure 4.20: Performance of CRM models for different sizes of input frame sequence. The frames were taken from videos (SEMAINE database) recorded at 50 fps.

**Effect of sequence length size in BLSTM**

In another set of experiments, the maximum length of the video sequence used as input to the BLSTM, was varied, to check its influence on the performance of the CRM$_{n=2}$ models. In this set of experiments, the videos in the SEMAINE database were clipped so as to form shorter sequences of length $N$ each (or less). As the video length may not be equally divisible by $N$, some video segments formed close to the end of the original videos, may have a length less than $N$. The models were trained and tested on the SEMAINE database for values of N = {25,50,100,500,3000}. The last value of 3000 corresponds to the maximum length of the original videos in the SEMAINE database. Fig. 4.21 shows the performance of AU models at different values of N. For most AUs, maximum performance was achieved when full length of the original videos was used (corresponding to N = 3000), with the exception of AU17 for which peak performance was achieved at N = 50. The average performance across all 6 AUs was also observed to be maximum when whole video sequences were used without any clipping. This was expected because clipping the videos to shorter sequences prevents learning of long range temporal information. Clipping the videos randomly to shorter length may also lead to

cropping out a part of the onset-peak-offset sequence of an AU which can be critical for learning the temporal evolution.

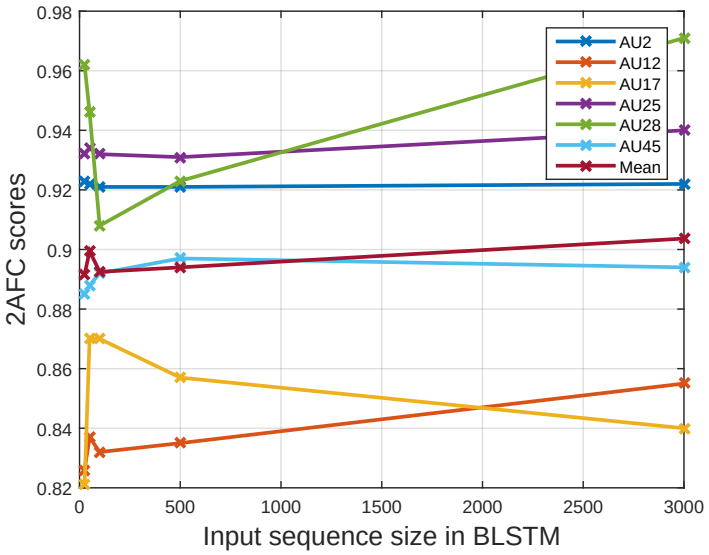


Figure 4.21: Performance of CRML models for different sizes of sequence length used as input to BLSTM.

**Performance comparison for occurrence detection**

In the next set of experiments, we compare the performance of our method ($\text{CRML}_{n=2}$) with other existing methods on the SEMAINE, BP4D and DISFA dataset. For SEMAINE and BP4D, we compare the performance of our approach with the Local Gabor Binary Patterns (LGBP, (Wu et al., 2012)), the SVM based approach described in (Valstar et al., 2015a)and the multi-label Discriminant Laplacian Embedding (DLE) approach proposed by Yüce et al.(Yüce et al., 2015) (FERA-2015 Challenge winner). Another method that we compare against is a geometric feature based approach which uses a deep neural network (GDNN). For computing the performance of this method we trained a deep neural network with 4 hidden layers (all fully connected). The input to this network were the locations of 49 facial landmarks within a time window of 5 frames. The 49 landmarks used in this case belong to the interior of the face. The 17 landmarks on the face contour were not used because these landmarks were assumed to be either not affected by any AU (landmarks near the ear) or correlated to some of the landmarks in the interior of the face (movement of contour landmarks on the chin can be correlated to the landmarks on the lower lip in case of AU25 and AU26). On DISFA database, we compared the performance of our approach against the CNN based framework of Ghosh

| AU | LGBP | GDNN | DLE | CRML$_{n=2}$ |
|---|---|---|---|---|
| 2 | 0.75 | 0.67 | 0.66 | **0.80** |
| 12 | 0.52 | 0.63 | **0.76** | 0.74 |
| 17 | 0.07 | 0.14 | 0.25 | **0.32** |
| 25 | 0.40 | 0.77 | 0.61 | **0.85** |
| 28 | 0.01 | 0.31 | 0.26 | **0.33** |
| 45 | 0.21 | 0.55 | 0.35 | **0.57** |
| Mean | 0.33 | 0.51 | 0.48 | **0.60** |

Table 4.3: Performance (F1 scores) comparison on SEMAINE test set. The proposed approach is compared against LGBP (Valstar et al., 2015c), GDNN and DLE (Yüce et al., 2015).

| AU | LGBP | GDNN | DLE | CRML$_{n=2}$ |
|---|---|---|---|---|
| 1 | 0.18 | **0.33** | 0.25 | 0.28 |
| 2 | 0.16 | 0.25 | 0.17 | **0.28** |
| 4 | 0.22 | 0.21 | 0.28 | **0.34** |
| 6 | 0.67 | 0.64 | **0.73** | 0.70 |
| 7 | 0.75 | **0.79** | 0.78 | 0.78 |
| 10 | 0.80 | 0.80 | 0.80 | **0.81** |
| 12 | **0.79** | 0.78 | 0.78 | 0.78 |
| 14 | 0.67 | 0.68 | 0.62 | **0.75** |
| 15 | 0.14 | 0.19 | **0.35** | 0.20 |
| 17 | 0.24 | 0.28 | **0.38** | 0.36 |
| 23 | 0.24 | 0.33 | **0.44** | 0.41 |
| Mean | 0.44 | 0.48 | 0.51 | **0.52** |

Table 4.4: Performance (F1 scores) comparison on BP4D Test set. The proposed approach is compared against LGBP (Valstar et al., 2015c), GDNN and DLE (Yüce et al., 2015).

et al. (2015), the approach of decision level feature fusion from facial regions (DFR) by Jiang et al. (2014a) and the incremental boosting CNN (IB-CNN) approach of Han et al. (2016). We used a 5 fold cross-validation for reporting the results on DISFA dataset.

Tables 4.3, 4.4 and 4.5 shows the performance comparison on the SEMAINE, BP4D and DISFA dataset respectively. In table 4.3, we can see that the performance from our approach is significantly higher on the SEMAINE dataset, as compared to other approaches. Similarly, we outperform the other approaches on the BP4D and DISFA dataset as well. Fig. 4.22 shows the weighted average performance on SEMAINE and BP4D dataset. The weights were calculated as the fraction of the number of frames in the database to the combined total number of frames in both databases. In this Fig. we can see that our method outperforms other methods (Yüce et al., 2015; Baltrusaitis et al., 2015; Gudi et al., 2015; Valstar et al., 2015a), on the FERA-2015 Challenge dataset.

**Performance comparison for intensity estimation:**

In the next set of experiments, we evaluate our approach on the task of AU intensity

| AU | CNN | DFR | IB-CNN | CRML$_{n=2}$ |
|---|---|---|---|---|
| 1 | 0.70 | 0.69 | **0.77** | 0.72 |
| 2 | 0.71 | 0.76 | 0.84 | **0.87** |
| 4 | 0.70 | 0.78 | **0.88** | 0.83 |
| 5 | 0.77 | 0.88 | 0.88 | **0.91** |
| 6 | 0.85 | **0.93** | 0.92 | 0.92 |
| 9 | 0.83 | 0.87 | 0.90 | **0.91** |
| 12 | 0.91 | 0.93 | 0.95 | **0.94** |
| 15 | 0.68 | **0.74** | 0.51 | 0.71 |
| 17 | 0.68 | 0.74 | 0.74 | **0.74** |
| 20 | 0.64 | 0.78 | 0.62 | **0.78** |
| 25 | 0.84 | 0.85 | 0.92 | **0.97** |
| 26 | 0.76 | 0.76 | 0.88 | **0.88** |
| Mean | 0.76 | 0.81 | 0.82 | **0.85** |

Table 4.5: Performance (2AFC scores) comparison on DISFA database for AU occurrence detection task. The proposed approach is compared against CNN based approach (Ghosh et al., 2015), DFR (Jiang et al., 2014a) and IB-CNN (Han et al., 2016)
.

estimation and compare our performance with other state-of-art methods reporting on the same databases. For this purpose, we used the BP4D and DISFA datasets and used all the AUs which are annotated with AU intensities. In BP4D dataset, only AU 6, 10, 12, 14 and 17 are annotated for intensities. We report the performance on this dataset for both the pre-segmented (only using examples which have intensity greater than zero) and unsegmented case. For BP4D pre-segmented dataset, we compare our performance against the approaches of LGBP (Valstar et al., 2015c), deep CNN (DCNN) described in (Gudi et al., 2015) and the person-specific normalization (PSN) approach described in (Baltrusaitis et al., 2015). In case of BP4D unsegmented case, we compare against LGBP(Valstar et al., 2015c), PSN(Baltrusaitis et al., 2015) and the hard multi-task metric learning approach (MLKR) described in (Nicolle et al., 2015). For DISFA dataset, we report our results on all 12 AUs using a 5-fold cross validation and also compare it with the Latent trees (LT) approach described by Kaltwang et al. (2015) and the Hidden Markov Model (HMM) approach described by Mavadati and Mahoor (2014).

Table 4.6 and 4.7 shows the performance on the BP4D test set for the pre-segmented and unsegmented case respectively. For the pre-segmented case, our method outperforms other methods in 4 out of 5 AUs. In case of unsegmented data, the average performance of our method is comparable to that of hard-MLKR (Nicolle et al., 2015) (Winner of the fully automatic intensity sub-challenge in FERA-2015). In both these tables, it can be observed that the performance for AU6 is significantly lower when compared to the best performing approach for this AU. A possible reason for this result is that it is difficult to

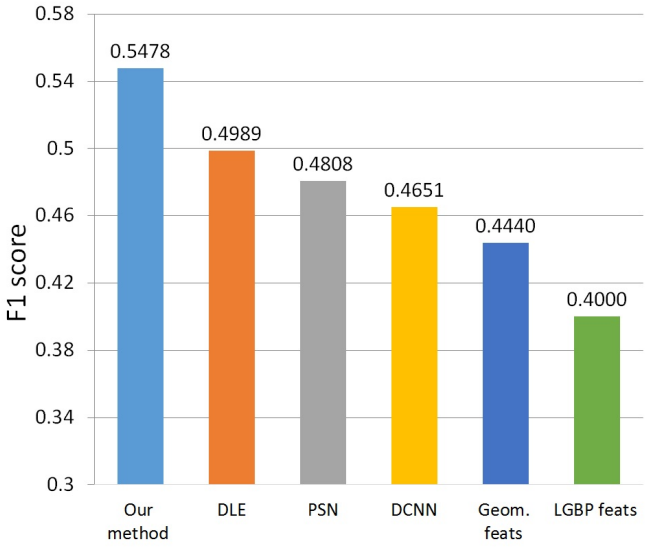


Figure 4.22: Weighted average performance on FERA-2015 test set (BP4D and SEMAINE) for AU occurrence. The weights were calculated as the fraction of the number of frames in the database to the combined total number of frames in both databases (SEMAINE+BP4D). Our method CRML$_{n=2}$ is compared against DLE (Yüce et al., 2015), PSN (Baltrusaitis et al., 2015), DCNN (Gudi et al., 2015) and Geometric and LGBP feats (Valstar et al., 2015c).

define the most optimum facial region for AU6 activation. Sub-optimal region definition for this AU could be one of the reasons for the lower performance. In addition, the problem of learning AU intensities is compounded by the fact that there could be large imbalance between different intensity levels in the training data. This could create bias in the learnt models thereby negatively affecting its performance. We did not specifically target this problem, which could have led to slightly lower performance as observed in this case. Table 4.8 shows the performance on the DISFA dataset. In this table it can be observed that our performance is higher as compared to the HMM approach of Mavadati and Mahoor (2014) and LT approach of Kaltwang et al. (2015).

| AU | LGBP | DCNN | PSN | CRML$_{n=2}$ |
|---|---|---|---|---|
| 6 | 0.33 | 0.42 | **0.60** | 0.43 |
| 10 | 0.48 | 0.54 | 0.53 | **0.54** |
| 12 | 0.60 | 0.61 | 0.73 | **0.73** |
| 14 | 0.50 | 0.49 | 0.49 | **0.61** |
| 17 | 0.10 | 0.22 | 0.47 | **0.56** |
| Mean | 0.40 | 0.46 | 0.56 | **0.57** |

Table 4.6: Performance (ICC scores) comparison on pre-segmented BP4D Test set for the task of AU intensity estimation. The proposed approach is compared against LGBP (Valstar et al., 2015a), DCNN (Gudi et al., 2015) and PSN (Baltrusaitis et al., 2015).

| AU | LGBP | PSN | MLKR | $CRML_{n=2}$ |
|---|---|---|---|---|
| 6 | 0.62 | 0.72 | **0.79** | 0.71 |
| 10 | 0.66 | 0.72 | **0.80** | 0.78 |
| 12 | 0.77 | 0.83 | 0.86 | **0.86** |
| 14 | 0.39 | 0.54 | 0.71 | **0.71** |
| 17 | 0.17 | 0.38 | 0.44 | **0.44** |
| Mean | 0.52 | 0.64 | **0.72** | 0.70 |

Table 4.7: Performance (ICC scores) comparison on unsegmented BP4D Test set for the task of AU intensity estimation. The proposed approach is compared against LGBP (Valstar et al., 2015a), PSN (Baltrusaitis et al., 2015) and MLKR (Nicolle et al., 2015).

| AU | HMM | LT | $CRML_{n=2}$ |
|---|---|---|---|
| 1 | 0.25 | **0.32** | 0.24 |
| 2 | 0.22 | **0.37** | 0.23 |
| 4 | 0.28 | 0.41 | **0.46** |
| 5 | 0.08 | 0.18 | **0.22** |
| 6 | 0.17 | 0.46 | **0.52** |
| 9 | 0.16 | 0.23 | **0.30** |
| 12 | 0.57 | 0.73 | **0.81** |
| 15 | 0.08 | 0.07 | **0.12** |
| 17 | 0.11 | 0.23 | **0.25** |
| 20 | 0.04 | 0.09 | **0.16** |
| 25 | 0.63 | 0.80 | **0.89** |
| 26 | 0.23 | 0.39 | **0.55** |
| Mean | 0.24 | 0.36 | **0.40** |

Table 4.8: Performance (ICC scores) comparison on DISFA database for AU intensity estimation task. The proposed approach is compared with HMM based approach (Mavadati and Mahoor, 2014) and Latent Tree based approach (Kaltwang et al., 2015).

### 4.2.4 Limitations and analysis of failure cases

We analysed the failure cases by observing the false positives returned by some of the AU occurrence detection models. Fig. 4.23, shows some typical false positives observed for AU2, AU4, AU12, AU15 and AU25. Taking a closer look at these examples, it can be observed that the problem cases for AU2 includes occlusion of the eye-brows due to presence of hair on the forehead, presence of glasses and out-of-plane rotation of the head. It can also be observed that the classifier returns false positives when the subjects in the images look upwards (eye pupil pointing upwards). A possible reason for this could be the fact that there are many AU2 examples in the training dataset where the subjects are looking upwards. The network seems to have learnt this correlation resulting in the observed misdetections.

For AU4, it can be observed that partially closed eyes are quite often detected as false positives. The correlation between these in the training data could be the reason for this observation. Non-frontal head pose is also observed in some of the failure cases for AU4.

In case of AU12, non frontal head pose again seems to be an issue. It can also be observed that visibility of teeth can lead to false positive detection of AU12. This is due to the high correlation between these facial events.

In case of AU15, the classifier seems to get confused with lip stretcher or even smiles wrongly classifying them as positives for AU15. Non frontal head pose is again a problem here which also sometimes result in incorrect facial landmark localization.

For AU25, it was observed that the classifier is very sensitive to facial landmark localization. This is because facial landmark locations on the lips are enough to tell if the mouth is open or closed. Therefore many of the failure cases observed for AU25 are due to incorrect localization of facial landmarks.

From the analysis of these failure cases it is clear that non-frontal head pose is currently one of the main limitation of the proposed AU detection model. Occlusion and presence of eye glasses can also cause misdetections. A number of false positives also occur due to the correlations between certain facial actions which don't always occur together (e.g. partially closed eyes with AU4, visibility of teeth with AU12, etc.). Another limitation is that for some AUs, the classifier can be quite sensitive to localization of the facial landmarks and consequently incorrect localization can lead to misdetections. It

remains to be seen that whether using larger datasets can alone overcome any of these limitations. Developing face frontalization techniques can be a good step in the future direction and can help overcome the head-pose problem without requiring additional AU data. Similarly, new techniques for handling occlusion and developing more reliable facial landmark trackers are good directions for future research. There are also some existing CNN visualization techniques such as visualization of the intermediate feature maps and filters learnt during the training (as mentioned by Breuer and Kimmel (2017)). These could be used to throw more insight into the workings of CNNs and can potentially suggest ways of improving the architecture itself. However, this is beyond the scope of this thesis and is considered a promising direction for future work.

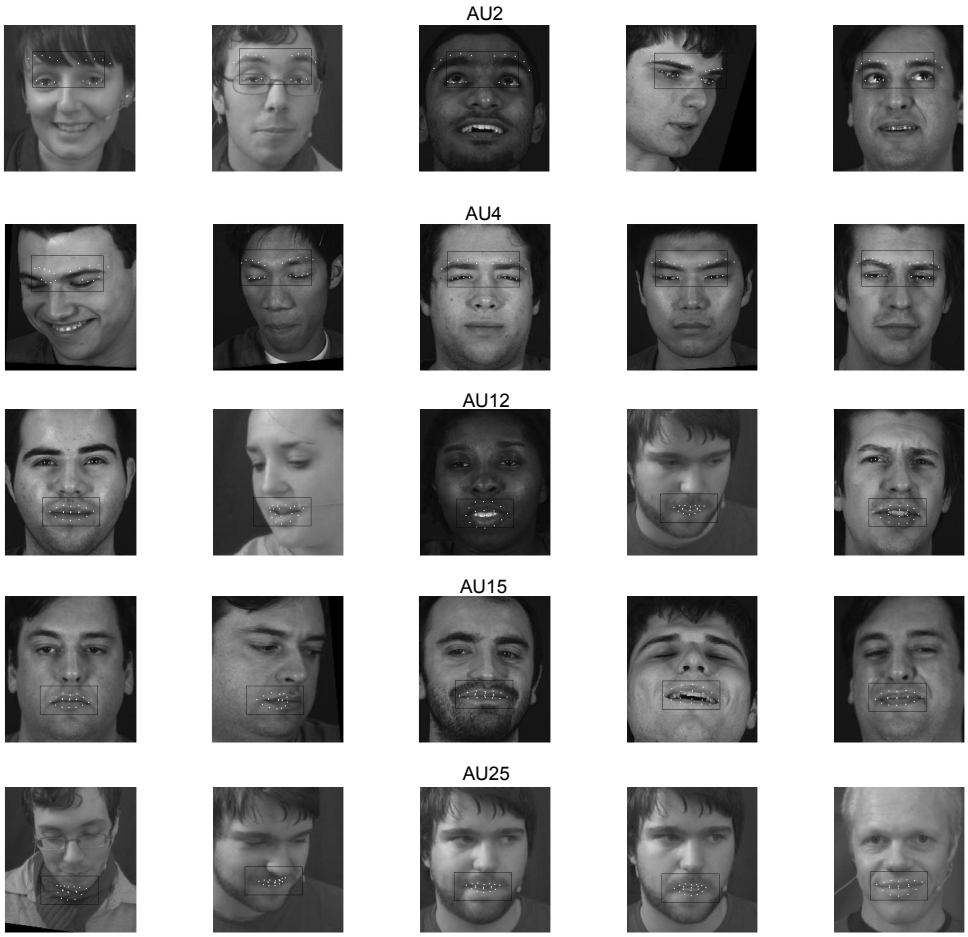


Figure 4.23: Examples of failure cases in AU detection task. Each row shows some example of the false positives observed for each AU. For each image, the locations of facial landmarks (denoted by white dots) and the facial region (denoted by black rectangle) used for the respective AU classifier, are also shown.

### 4.2.5 Computational cost analysis:

Finally, we also analysed the computational cost of the main components of the proposed system. Table 4.9 shows the breakup of the computation time and internal memory requirement of an AU model at the evaluation stage. These costs have been calculated separately for the pre-processing, CNN and BLSTM stage on a video consisting of 100 frames. The costs have been computed for an Intel Core i7 CPU. Here the pre-processing stage includes face detection, facial landmark detection, face registration, facial region cropping and binary mask computation. In Table 4.9, it can be observed that in terms of computation time requirement, the major current bottleneck is the pre-processing step. The BLSTM evaluation on the other hand takes only 0.32 sec. However, it should be noted that the implementation used for BLSTM was written in C++ while for the pre-processing steps and CNN evaluation, an unoptimized MATLAB code was used. In terms of the internal memory requirement, it can be observed that BLSTM evaluation requires the largest amount of memory. Moreover, the memory required by BLSTM increases linearly with the length of the sequence (number of frames) as the entire sequence need to be kept in memory. This additional(variable) memory requirement for BLSTM can become a significant problem for processing videos of long duration.

| | Pre-processing | CNN | BLSTM |
|---|---|---|---|
| Time requirement (in sec.) | 28.53 | 6.96 | 0.32 |
| Memory requirement (in MB) | 400 | 520 | 711 + 2.4(Variable) |

Table 4.9: Computational cost of different components of the AU detection system. The cost is calculated for applying an AU model on video consisting of 100 frames. The variable part (in BLSTM) is the additional memory required by the sequence of input features (depending on sequence length).

## 4.3 Conclusion

This chapter described a novel CNN-BLSTM based approach which learns the dynamic appearance and shape of facial regions for facial expression recognition. The appearance and shape are learnt through local image regions and corresponding binary masks respectively. The dynamics are learnt through a combination of dynamic features (extracted from a time-windowed CNN) and BLSTM. It was shown that each component of the proposed system contributes towards an improvement in performance and achieves

performance which is comparable or higher than the current state-of-the-art on a number of databases. The approach was evaluated for facial AU detection and AU intensity estimation tasks.

# Chapter 5

# A study on ADHD and ASD patients for visual data recording

Evaluation of any medical diagnostic tool relies on the availability of large databases. Machine learning based methods especially require annotated databases which can be used for both training and testing purposes. Hence databases are a critical component of research in this field. A major challenge of working in the area of medical and mental health technology is that it is hard to find publicly available databases which can be used for evaluating new approaches. Due to the sensitive nature of the data, researchers find it difficult to get ethical approval for sharing such databases with other fellow researchers. Even if a database is available which can be shared, it might not be very suitable for evaluating a specific kind of approach. In such cases, a new database needs to be created which is recorded by taking into account the specific needs of the proposed approach.

One of the goals of this thesis was to develop a method which uses automatic facial behaviour analysis to help in an objective diagnosis of ADHD and ASD. Evaluation of such an approach would require a database consisting of visual data not only from people with ADHD and ASD, but also from healthy controls who do not show any symptoms of these conditions. The database should be suitable for accurate automatic facial expression recognition and 3D head-pose estimation. For this purpose, the database would need to be recorded in a controlled setting where the facial images can be captured in near frontal positions. Availability of depth data along with colour images would also enable accurate 3D head-pose estimation. Above all, there should be sufficient number of people in each possible group (ADHD, ASD, Controls, etc.), to measure any statistical

difference in their facial behaviour.

This chapter describes a new **K**inect database for **o**bjective **m**easurement of **A**DHD and **A**SD (KOMAA). This database consists of RGBD (Colour + depth) video recordings of clinical and control participants recorded using a Microsoft Kinect sensor. The database was recorded to evaluate the approach proposed in this thesis for automatic detection of people with ADHD and ASD using facial behavioural features (described in Chapter 6). This chapter gives an overview of the KOMAA database including recruitment of participants, screening questionnaires used for ADHD & ASD, recording of videos using the Kinect sensor camera and details of the computer based tasks which the participants were asked to do during the recording. A statistical overview of the participants in the recorded database is also given.

## 5.1 Participant recruitment

Clinical participants were recruited with the help of the Nottingham Asperger service and ADHD clinic situated in Nottingham (UK). All patients undergoing ADHD and ASD assessment were invited to take part, providing they do not have an intellectual disability, are over the age of 18 and can provide informed written consent. Healthy controls were recruited from around the University of Nottingham, via generic e-mail lists and poster advertisements displayed on appropriate notice boards. The recruited participants were either students or employees at the University of Nottingham. The recruited control participants were also required to be over the age of 18 and capable of providing informed written consent. The ethics approval for this recruitment and the corresponding study was obtained from the NHS research ethics committee (NRES-The Black Country, REC ref: 15/WM/0161, Date of approval: 30/06/2015).

To avoid any overlap between the clinical and healthy controls, all participants in the healthy control group had to score below a certain threshold value on the screening measures for ADHD and ASD. The following are the two screening measures used in this study:

1. **Autism Quotient AQ10:** This widely used screening measure for Autism symptoms is a questionnaire consisting of 10 items related to the symptoms of Autism.

The measure covers the following 5 domains relevant to Autism: i) attention to detail ii) attention switching iii) communication iv) imagination and v) social. The measure has excellent reliability and validity. The scores from this questionnaire ranges from 0 to 10. People scoring 6 or more on this questionnaire are usually advised to go for a comprehensive assessment of ASD. See Chapter 2 for more details. A full copy of this questionnaire can be found in Appendix A.

2. **Adult ADHD Self-Report Scale (ASRS-v1.1):** This is a screening measure for ADHD which was developed by the World Health Organization (WHO). The eighteen item screening measure for ADHD symptoms in adulthood is based on the DSM-IV items for ADHD. It is divided into 2 parts (Part-A and Part-B). Part-A has 6 questions which are considered to be highly indicative of ADHD. Part-B has 12 other questions related to the symptoms of ADHD. People scoring 4 or more in Part-A of the questionnaire are usually advised to go further for a comprehensive evaluation. The total score from the entire questionnaire ranges from 0 to 18, with point possible from each question. A full copy of this questionnaire can be found in Appendix A.

The participants from both the control group and the clinical group were asked to complete the above screening questionnaires. For the control group, only those participants who scored less than 6 in AQ10 and less than 4 in ASRS (Part-A), were included in the database.

## 5.2 Recording of data

All participants were invited to participate in the study at the Nottingham Asperger service and ADHD clinic or at the School of Computer Science , University of Nottingham. All control participants participated in the study at the School of Computer Science, University of Nottingham. All clinical participants participated in the study at the Nottingham Asperger service, except two participants who preferred to come to the University.

During the recordings the participants sit in front of a computer screen and have to read and listen to a set of 12 short stories selected from the "Strange Stories" task (Happé, 1994). Each story is described in text and depicted by a picture. It consists of situations involving people saying something non-literal which may include a Lie, White lie, Joke,

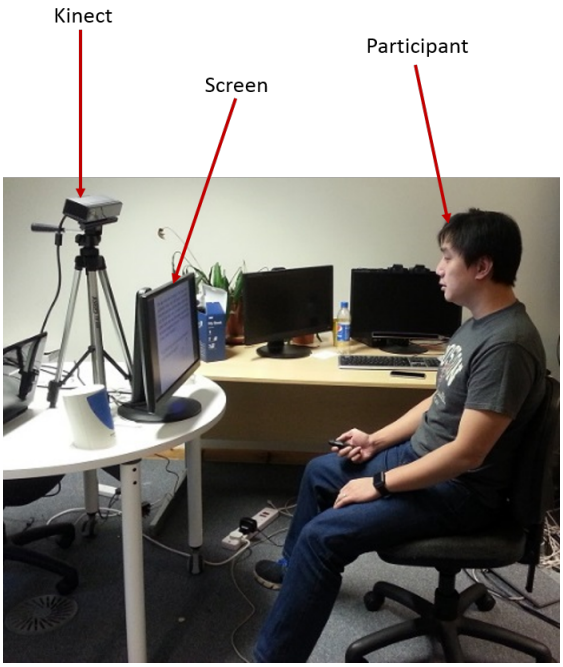


Figure 5.1: Recording setup.

Pretend, Misunderstanding, Persuade, Appearance/Reality, Figure of Speech, Sarcasm, Forget, Double Bluff, and Contrary Emotions Happé (1994). Participants were asked to answer 2-3 questions about the intention of the character described in each vignette (a copy of the 12 Strange Stories used in this study can be found in Appendix B). This task was originally designed to measure the mentalizing abilities of a person. For this study, a computer version of this task was created in which the participant can read the stories on a computer screen. Additionally, a pre-recorded voice was played reading out the story and the corresponding questions. Participant were required to answer the questions verbally. Such a setup was designed so as to simulate the effect of a real person to person conversation, while at the same time keeping the setup as automated as possible. During the entire task, the participant's RGBD video and the corresponding audio data is recorded using the Kinect device. The Kinect device is placed directly behind the computer screen so as to capture the frontal view of the participant (See Fig. 5.1 for an overview of the recording setup).

The data was initially recorded using the Kinect Studio software from Microsoft, which is specially designed for Kinect. However, the files created by this tool are very large in size (a minute of recording produces files of size over 8GB). This is because the video data in these files are stored in an uncompressed format. For each session, the resulting files were of size close to 100GB. In order to compress the data and also to extract the

individual data streams (depth maps, colour video) and other meta-data (face tracking data, skeleton tracking data, etc.), the Social Signal Interpretation (SSI) framework developed by Wagner et al. (2013), was used. The final database consists of 6 files for each session. These include the RGB video (in MPEG-4 format), depth data, 3D facial landmarks data, 3D head-pose data, body skeleton tracking data and the Kinect facial Animation Units data (see Chapter 6 for more details on Animation Units). Except the RGB video, all other files are stored in binary format. The total size of files for each session is approximately 4GB. All the data files were anonymised by removing any personal details like names etc. and replacing them with an identity number.

Apart from the Strange Stories task, the participants were asked to complete two additional tests which are often used as a diagnostic tool for the assessment of ASD and ADHD. The purpose of conducting these additional tests was to get an estimate of how well the scores from such tests agree with the actual diagnosis by the clinicians and with the symptom scores from questionnaires like AQ10 ans ASRS. The following additional tests were conducted during the study:

- **Reading the Mind in the Eyes test (RME)** (Baron-Cohen et al., 1997, 2001a): In this test, participants are shown pictures of human faces each having 4 words around it. The participants are asked to pick out the word which best matches the emotion depicted by the person in the image. The test is designed to assess the ability of the participants to comprehend emotional states from facial gestures and expressions. This ability is often impaired in adults with ASD and therefore this test is frequently used in the diagnosis of ASD. The whole test consists of a total of 37 images (including one for practice).

- **QbTest:** This is a computer based test in which a participant is required to respond quickly and accurately to certain geometric shapes displayed on the screen. The purpose of this task is to measure inattention and impulsivity which is often observed in adults with ADHD. The test lasts for 20 minutes and during the entire duration, the participant's head movement is tracked through a reflector and an infra-red camera (see Chapter 2 for more details). At the end of the test, the participant's head tracking data and the performance on the computer based task, are compared against the norm data to generate scores indicating the participant's activity level, attention level and impulsivity during the test.

## 5.3 Overview of the recorded dataset

The subjects in this database can be divided into four different categories. The first category is the control group which consists of subjects who show no symptoms of ADHD/ASD and have never been diagnosed with either ADHD or ASD. The other three categories include the ASD group (consisting of subjects who have been diagnosed with ASD), ADHD group (subjects who have been diagnosed with ADHD) and ASD+ADHD group (subjects who have been diagnosed with both ADHD and ASD). The total number of subjects recruited into each category is shown in Fig. 5.2. In this Fig. it can be observed that the number of participants in the ADHD category is much lower than in other categories which have at least 10 or more participants each.

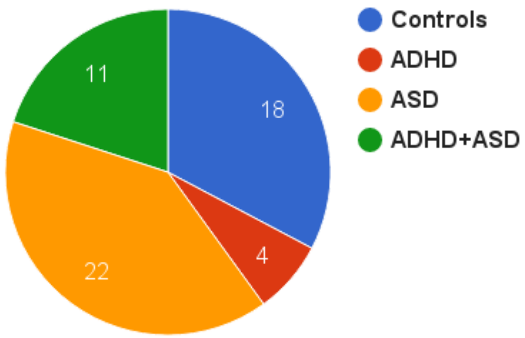


Figure 5.2: Distribution of participants in KOMAA dataset.

The gender and age distribution of the participants in each category can be seen in Fig. 5.3 and 5.4 respectively. In terms of gender distribution, the dataset is unfortunately imbalanced. While the number of females is higher than the number of males in the control group, in all other groups the number of males is much higher. As far as the age distribution is concerned, the median ages of the participants in the ASD and Comorbid (ASD+ADHD) are slightly higher than that of the control and ADHD groups.

In order to test how well the symptom scores provided by AQ10 and ASRS questionnaires agree with the clinical diagnosis, the distribution of these scores were plotted for all 4 types of diagnosis. Fig. 5.5 shows the mean AQ10 and ASRS scores and their standard deviation (denoted by the error bars) for each group of participants. In this Fig. it can be observed that the participants in the control group score significantly lower than the other groups, which adds confidence to the assumption that the participants in the control group do not show any symptoms of ADHD or ASD and hence can be safely regarded

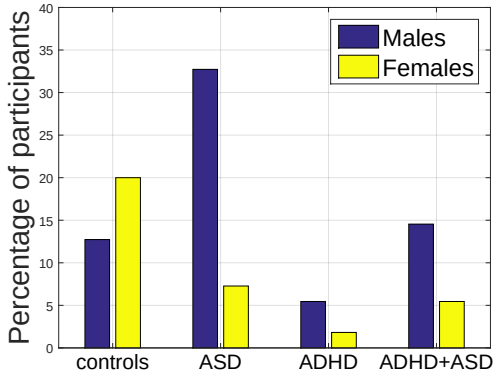


Figure 5.3: Gender distribution of participants in KOMAA dataset.

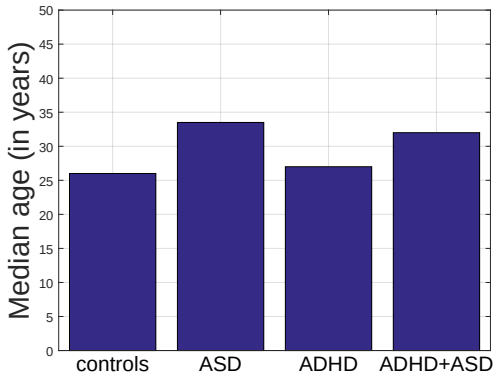


Figure 5.4: Median age of participants within different groups in the KOMAA dataset.

as "controls" for any further analysis. Regarding the clinical groups, on average it can be observed that the AQ10 scores are lower in the ADHD group as compared to the ASD and Comorbid (ASD+ADHD) groups. However, the difference is not as high as the difference between the control and all other clinical groups. A similar pattern can be observed for the ASRS scores which on average are lower in the ASD group as compared to ADHD and Comorbid group. However, here again the difference is not as sharp as the difference between the control group and the other clinical groups. These observations indicate that there is an overlap between the symptoms of ADHD and ASD captured by the AQ10 and ASRS questionnaires. It should be noted that this study was explorative in nature and due to the small number of samples in each category, the results described here are only for indication.

Finally, the distribution of the scores from another independent test ("Reading the mind

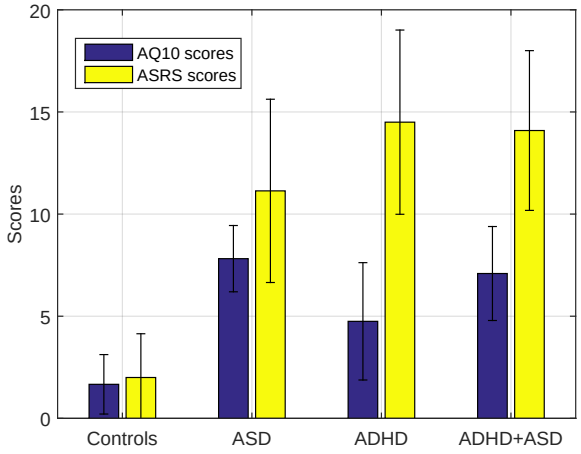


Figure 5.5: Distribution of average AQ10 and ASRS scores within different groups.

in the eyes") was also plotted for the controls and each clinical group. Fig. 5.6 shows the mean scores from Reading the mind in the eyes (RME) test, for each group of participants. The standard deviation in their scores is denoted by the respective error bars. For the RME test, it can be observed that the average score in the control group is only marginally higher than the ASD group. Also, the scores in the ADHD group are actually lower than the ASD group. This shows that the RME test does not help significantly (at least for this dataset) in discriminating between people with and without ASD.

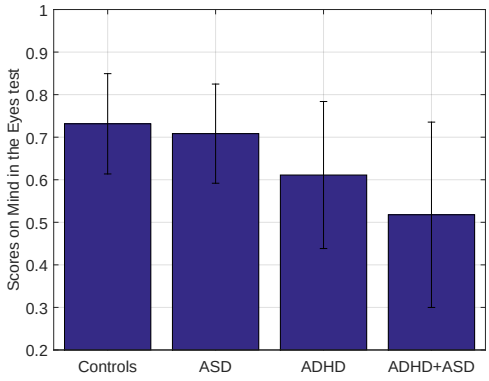


Figure 5.6: Distribution of "'Reading the Mind in the Eyes"' test scores within different groups of participants.

## 5.4 Conclusion

This Chapter described a new database consisting of audio and visual (RGBD) data from healthy control participants and from clinical participants who have been diagnosed with either ADHD or ASD or both (comorbid). The database can be used to evaluate methods which utilize human behaviour analysis for automatic prediction ADHD and ASD. The recording setup was specially designed to simulate a real person to person conversation but at the same time keeping the scenario controlled and the recording setup as automated as possible. However, the current setting can still be improved to more closely match the way humans communicate with each other. In the current setup, the whole text of each story is displayed on the the screen from the beginning while an audio narration of the same story is played in the background. This could be replaced by display of only the current text/sentence which is being narrated by the background voice. Using virtual humans instead of displaying text may also make the conversation more realistic.

Apart from the audio-visual data, the database also has information about each participant's scores on screening measures like AQ10 and ASRS, scores on an independent test like "Reading the mind in the eyes" and their diagnosis by the clinicians. A preliminary analysis of the data showed that using existing tools, it is relatively hard to differentiate between ASD, ADHD and Comorbid conditions as compared to discriminating between healthy controls and people with either of these conditions. This is due to the overlapping of symptoms captured by the existing measures. Human behaviour analysis especially facial gestures can provide an alternate source of information could possibly help in differentiating between these conditions. An approach targeted in this direction is described in Chapter 6.

# Chapter 6

# Automatic Detection of ADHD and ASD from Expressive Behaviour

In Chapter 2, we learnt about neurodevelopmental conditions like ADHD and ASD which affect a significant part of the population. We also learnt about the current methods of their diagnosis which includes screening questionnaires, diagnostic interviews and looking for certain behavioural markers through manual observation. Such methods for their diagnosis are not only costly, difficult to repeat and time consuming but are also susceptible to human decision making bias. In Chapter 5, it was also observed that the existing diagnostic tools like the screen questionnaires (e.g. ASRS and AQ10), "Reading the mind in the eyes" test and Qb test are not sufficient for distinguishing these conditions. This is due to the overlap of symptoms captured by such tools. On the other hand, it is known that these conditions alter expressive behaviour which if analysed automatically, can help in a more efficient and objective diagnosis. This kind of approach was formalized by Valstar (2014) as *Behaviomedics*. Automatic analysis of expressive behaviour has seen a steady progress in the past few years and the current computer vision based algorithms can perform reliable face tracking and facial expression recognition, at least under mild environmental constraints. In Chapter 4, we presented our own deep learning based approach for facial Action Unit detection and intensity estimation, achieving state-of-the-art performance on a number of databases. Building a system which can harness the capabilities of such algorithms to aid diagnosis of ADHD and ASD, can help

in bringing more objective, repeatable measures in the decision making process.

This Chapter describes a novel approach which aims to make the diagnostic procedure for ADHD and ASD easier, more efficient and more objective through automatic analysis of a person's facial behaviour. In this chapter, a machine-learning based approach is proposed, to automatically aid diagnosis of ADHD and ASD. This approach involves extraction of high level features from tracked faces in videos to learn classification models for ADHD and ASD prediction. It uses a version of the method proposed in Chapter 4, for facial AU recognition, and face tracking data from RGBD (colour+depth) images recorded using a Kinect 2.0 sensor camera to obtain head actions and facial animation unit parameters. The proposed approach is evaluated on the KOMAA database described in Chapter 5. This database was collected specifically for this task in which 55 subjects who have previously been diagnosed with ADHD or ASD as well as subjects from a healthy control group were recorded in a controlled setting (please refer to Chapter 5 for more details). There are no existing publicly available databases suitable for testing the proposed method. Hence this methodology has been evaluated only on the KOMAA dataset.

## 6.1 Methodology

Training statistical machine learning based classifiers which can automatically differentiate between subjects with ADHD/ASD from healthy controls, is a difficult problem. The problem becomes even more challenging when the number of training examples is small. Deep learning based approaches which directly use low level pixel information to learn high level semantics, currently provide state-of-the-art performance on a number of computer vision tasks. However, using low level information on the limited number of training examples in our case, can lead to severe overfitting.

Our approach to training the classifiers involves computing high level feature descriptors corresponding to facial expressions (facial AUs), head pose and head motion. To compute the behaviour descriptors, each video is first divided into 12 segments corresponding to the 12 stories that the participants have to read while they were recorded. This has been done manually, but could easily be automated given that the timing of the delivery of the stories is controlled by the researcher.

For each video segment, histogram based feature descriptors are computed separately using pre-trained classifiers/regressors that detect individual behavioural cues. Grouping these cues per story helps to preserve temporal information as well as context specific facial behaviour in response to each story, which would otherwise be lost if histograms would have been computed over all the frames in a video, at a small price of multiplying the dimensionality of our overall feature vector by a factor 12. The combined set of feature descriptors from all segments in a recording are used for used for training the ADHD/ASD classification models (See Fig. 6.3). Below we describe the main components of our approach in more detail.

### 6.1.1 Feature descriptors

Six different sets of features are computed from the recorded video of each subject. Most of the features are computed on a per-frame basis, which are then converted into multiple histograms where each histogram is computed over all the frames in a video segment. The feature descriptors used in our approach are described below:

**1) Dynamic Deep Learned Facial Action Units:**
Facial Action Units (AU) are movements of individual or groups of facial muscles defined according to the Facial Action Coding System (FACS) (Ekman et al., 2002). Anatomically based descriptors of facial expressions, they can be representative of the emotional and mental state of a person and can encode a large number of social signals. Intensities for a set of 6 AUs (AU1, AU2, AU4, AU12, AU15, AU20) and occurrence for AU45 (blinks) were estimated for each frame in video. For this purpose, we used AU intensity models trained using a version ($\text{CRM}_{n=2}$) of the deep CNN based approach described in Chapter 4. Due to larger memory requirement for BLSTMs (which increases linearly with the length of the input video) compared to the performance gain obtained ( 1%), the AU models used here do not employ BLSTMs (See Table 4.9 in Chapter 4 for more details on the computational cost requirement). The models were trained on the DISFA dataset. The network architecture used for this purpose is shown in Fig. 6.1. Histograms of AU intensities are computed over all the frames in a video segment. One histogram $H_{au_i}$ was computed for each AU consisting of 10 equally spaced bins each, covering the entire range of intensities (0 to 1). For AU45, the frequency of its occurrence and the

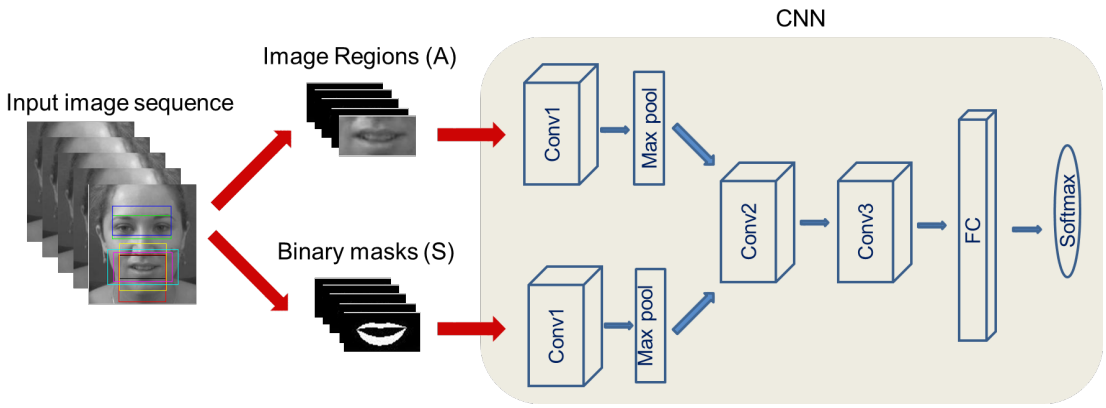


Figure 6.1: Graphical overview of the CNN based approach used for predicting facial AU intensities.

average duration of its activation are estimated in each video segment. The histograms of all AU intensities and the AU45 statistics $S_{au45}$ were concatenated together resulting in a 62 dimensional AU vector $F_{au}$, for each video segment.

$$F_{au} = [H_{au1}\ H_{au2}\ H_{au4}\ H_{au12}\ H_{au15}\ H_{au20}\ S_{au45}] \tag{6.1}$$

**2) Kinect Animation Units:**

The Kinect also provides Animation Units (AnUs), geometry-based descriptors similar to mpeg-4 face animation parameters (FAPs) (Pandzic and Forchheimer, 2003). While they are not based on muscle actions and can not detect facial actions that only cause appearance changes, the fact that they are obtained from RGBD data and computed in real-time by the Kinect library, it has been utilized by a number of facial expression recognition systems (Alabbasi et al., 2015; Mao et al., 2015). The intensity of a number of AnUs were estimated for each frame in the video using the Kinect v2 library. In order to aggregate the statistics over each video segment, a histogram of ANU intensities was computed for each facial AnU. Each histogram consisted of 10 bins resulting in a 10 dimensional feature vector corresponding to each ANU. A total of 10 AnUs (5 corresponding to left and 5 to right part of the face) were used. We used AnUs corresponding to lip stetcher (AnU4, AnU5), lip corner puller (AnU6, AnU7), lip corner depressor (AnU8, AnU9), eye closed (AnU12, AnU13) and brow lowerer (AnU14, AnU15). The histograms from all 10 AnUs were concatenated, resulting in a single AnU feature vector $F_{an}$, for each video segment.

**3) Head Pose:**

One of the major challenges for people with ADHD is their inability to do tasks which require sustained attention. The pose of the head (in 3D space) can provide valuable cues about the attention state of a person at a certain instance of time. Since the participants in our study were required to complete the task by looking the computer screen, any deviation of the head pose away from the computer screen would indicate loss of attention.

The rotation of the head about the X, Y and Z axis (pitch, yaw and roll) were estimated for each frame of the video using the Kinect 2.0 software (See Fig. 6.2). The X, Y and Z axis are defined in reference to the location of the Kinect device as shown in Fig. 6.2. We assumed the median pose of the head to be the most attentive state. Rotation of the head away from the median pose were computed about the X, Y and Z axis separately. Rotation in the negative and positive direction (for each axis) were assumed to be equivalent. Histograms of these rotation angles ($H_{RotX}$, $H_{RotY}$, $H_{RotZ}$) were computed over the video segments for each axis separately. Each histogram consisted of 9 bins with equally spaced bin centres ranging from $0^{\circ}$ to $45^{\circ}$. These histograms were concatenated resulting in a 27 dimensional head pose vector $F_{hp}$, for each video segment.

$$F_{hp} = [H_{RotX}\ H_{RotY}\ H_{RotZ}] \tag{6.2}$$

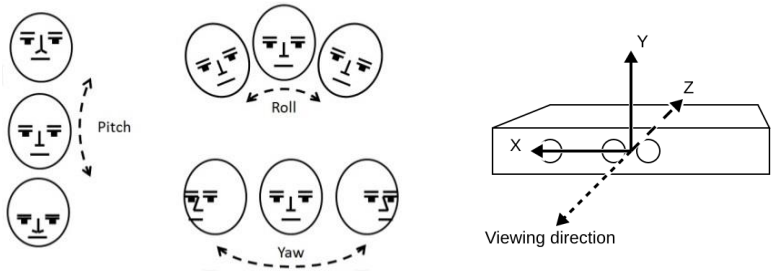


Figure 6.2: Rotation of head about X, Y and Z axis defined according to the Kinect coordinate system. Images taken from *https://msdn.microsoft.com*

**4) Speed of head movement:**

Dynamics of head motion has been a less researched aspect in the field of psychological disorders. In order to investigate the role of head motion, we estimated the speed of head motion at each frame of the video. For this purpose, we selected a set of stable facial landmarks (obtained from Kinect) belonging to eye corners and 4 points on the nose. The location of these stable facial landmarks are invariant to changes in facial

expressions and hence suitable for estimating the motion of the head. The motion of the head is estimated by computing the location of the centroid $C_i = \frac{1}{N}\sum_{j=1}^{N} X_{ij}$ of the stable landmarks $X_{ij}$. The speed of head motion $S_i$ at any frame $i$ can be estimated by computing the displacement of the centroid as given below:

$$S_i = ||C_i - C_{i-1}|| * f \tag{6.3}$$

where $f$ is the frame rate of the recorded video. In order to make speed estimation more reliable and invariant to any fluctuations in the frame rate, the estimated speed was smoothed by computing a moving average over 20 consecutive frames.

A histogram of the estimated speeds was computed to aggregate the statistics over each video segment. The histograms consists of 10 bins resulting in a 10 dimensional speed vector $F_{sp}$, for each video segment.

**5) Cumulative Distance:**

Hyperactivity is another major challenge associated with ADHD, implying that individuals with ADHD tend to display much higher levels of motoric behaviour than healthy individuals. The movement can be in the form of whole body movement or smaller movements confined to head (rotation) or hands and legs (fidgeting). To encode such information, the cumulative distance $F_{cd}$ moved by the head during an entire video segment, was estimated by summing up the displacements of the centroid $C_i$ (computed from facial landmarks obtained from Kinect) given below:

$$F_{cd} = \sum_{i=1}^{n} ||C_i - C_{i-1}|| \tag{6.4}$$

where n is the total number of frames in the video segment.

**6) Response Times:**

The time taken to respond to each set of questions in the study was also used as features. Since there were 12 stories, each comprising a set of questions, a 12 dimensional response time vector $F_{rt}$ was defined consisting of the response times (in seconds) for each set of questions.

### 6.1.2 Feature pre-processing and training models

**Normalization:** Each set of features (except the $F_{rt}$) were divided by the total number of frames in the video segment, to make them invariant to the length of video recording. In order to encode facial behaviour statistics from the entire video, a final set of features $F$ was obtained by concatenating all sets of features ($F_{au}$, $F_{an}$, $F_{hp}$, $F_{sp}$, $F_{cd}$, $F_{rt}$) from all video segments. Here the feature set from each video segment contributes in encoding the context specific facial behaviour in response to a specific story in the Strange Stories task. Each dimension in the resulting feature vector $F$ is further normalized by computing the Z-scores given by:

$$Z_{ij} = \frac{F_{ij} - \mu_i}{\sigma_i} \tag{6.5}$$

where $F_{ij}$ is the $i^{th}$ feature in the $j^{th}$ example, $\mu_i$ is the mean and $\sigma_i$ is the standard-deviation of all values in the $i^{th}$ feature dimension.

**Feature selection and training models** Due to the high dimensionality of the resulting feature $F$ compared to the number of training examples, any classifier trained directly on the entire feature-set is most likely to overfit the training data. In order to avoid this problem, a greedy forward feature selection was employed to capture the most relevant features and reduce the dimensionality. This feature selection method was preferred over other dimensionality reduction techniques (e.g. Principle Component Analysis (PCA), Correlation feature Selection (CFS)) because it can be used to find an optimum feature set by directly optimising on the prediction accuracy, in contrast to other methods like PCA or CFS in which the objective function is not directly related to the performance metric. The classification models were trained using Support Vector Machines (SVM) with a Radial Basis function kernel. SVMs have been one of the most widely used classifiers due to their convex optimization problem (no local minima) and their efficient use of kernels which gives them the ability to model non-linear problems. SVMs also more suitable for problems involving smaller number of training examples as compared to deep neural networks (state-of-the-art) which usually require large amount of training data.

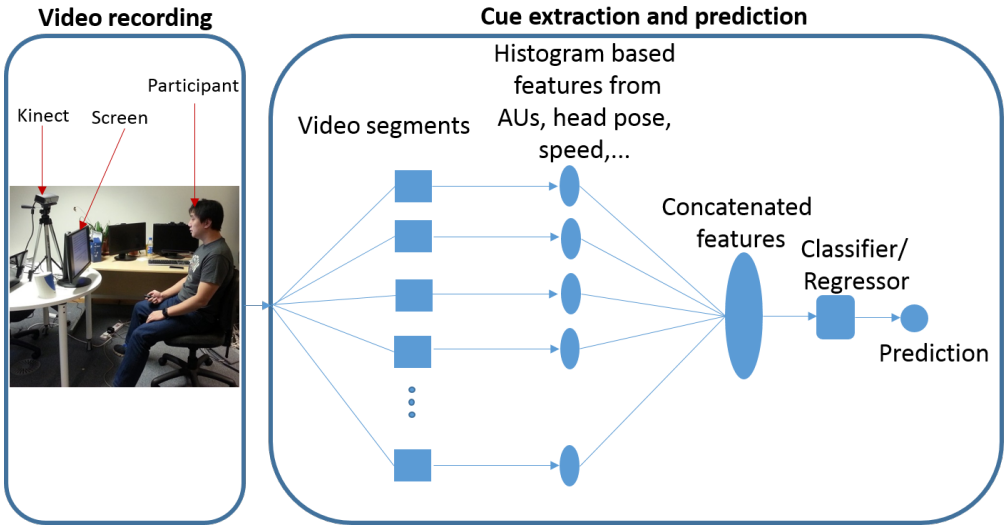


Figure 6.3: Overview of our system. A participant follows instructions on a screen while being recorded by a Kinect 2 camera. Deep Learning and RGB-D behaviour analysis of each video segment leads to successful ASD/ADHD classification.

## 6.2 Experiments

A number of experiments were conducted in order to evaluate the proposed approach. All the experiments were conducted using the KOMAA database described in Chapter 5. The performance of the proposed algorithm was measured for classification of each subject into ASD, Comorbid (ADHD+ASD) and Control group. In addition the class separation provided by the features and the distribution of selected features among different groups, was also analysed. It should be noted that since the ADHD only group was too small (only 4 participants), there was not enough data to learn a robust classifier for this group.

### 6.2.1 Performance evaluation

To evaluate the performance in classifying each subject to the ASD, Comorbid or the Control group, a 2 step procedure was followed: In the first step a classifier was trained to distinguish between control and condition groups (participants diagnosed with either ADHD, ASD or both). In the second step, another classifier was trained to distinguish between the ASD group and Comorbid (ASD+ADHD) group. When training both kinds of classifiers, the feature selection step was done only once on the entire dataset. This

was necessary to obtain a reliable set of features from a small sized dataset. The final performance was however evaluated using a leave-one-subject-out protocol, in which one subject is used for testing and the rest of the subjects are used for training the classifiers. This process is repeated for each subject and the overall score is obtained by averaging over each test subject. The classification performance is shown in Table 6.1 and 6.2. For classification into Control and Condition group, a very high classification accuracy of 96% was obtained. Similarly, for classification into Comorbid(ASD+ADHD) and ASD only group, a classification accuracy of 94% was obtained. It should be noted that if the feature selection is performed only on the training set (using an inner 10 fold CV), the classification accuracy drops to 87% for Controls vs Condition classification and to 55% for ASD vs Comorbid classification, when using the forward feature selection approach. We also experimented with the GentleBoost (Friedman et al., 2000) approach for feature selection (using only the training set). This resulted in a classification accuracy of 89% for Controls vs Condition classification and 73% for ASD vs Comorbid classification. A possible reason for the large drop for ASD vs Comorbid classification rate could be the relatively low number (33) of training examples available. For any further analysis below, only the results obtained from the forward feature selection (over the entire dataset) will be used.

Looking closely at the two incorrectly classified subjects for Control vs Condition classification, it was observed that both the subjects appear to be showing slightly more activity (in terms of head motion) which resulted in a speed profile looking similar to the condition group. Additionally, it was noticed that one of the subject is also wearing glasses which could have prevented reliable detection of AU1 intensity. In the ASD vs Comorbid group classification category, it was observed that for one of the incorrectly classified subject Kinect face tracking failed for the first 25% of the entire video duration. This would have resulted in default (zero) values to be used for the features corresponding to the untracked part and consequently may have lead to its misclassification. The other incorrectly classified subject in this category was also found to be wearing glasses which make the estimation of AU1 intensity difficult.

Table 6.1: Classification results for Controls vs Condition (ASD/ADHD) group.

| Classifier | Correct | Incorrect |
|---|---|---|
| Controls | 16 | 2 |
| Condition | 37 | 0 |

Table 6.2: Classification results for Comorbid (ADHD+ASD) vs ASD group.

| Classifier | Correct | Incorrect |
|---|---|---|
| Comorbid | 9 | 2 |
| ASD only | 22 | 0 |

Looking at the individual contribution of different cues, Fig. 6.4 and 6.5, shows the class separation provided by some of the important features selected by using the forward feature selection approach. From these figures, it can be observed that for classification of Controls and Condition group, features such as Speed of head motion (from video segment corresponding to story 1 and 2) and Animation Unit 8 (lip-corner depressor from video segment corresponding to story 10 of 'Stange Stories task') were found to be most discriminative. For Comorbid vs ASD classification, AU1 (inner-brow raiser), AnU6 (lip-corner puller) and head rotation about Y-axis turn out be highly discriminative These features were extracted from the video segment corresponding to story 1, 3 and 8 of the Strange stories task respectively. From a rough visual analysis of the participant's videos, it can also be observed that activity levels, head and eyebrow movements can be good discriminative features for the classification of these conditions. The results in Fig. 6.4 and 6.5 confirm this hypothesis. Fig. 6.6 and Fig. 6.7 also shows a list of top 30 features (for both classification problems) ranked according to their individual classification power.

The above results indicate that the facial behaviour of people with ADHD/ASD can be different compared to people without these conditions. This difference can be exploited for the diagnosis of these conditions. Differences in terms of activity levels and inattention (for people with ADHD) is already well known. However, the above results indicate that are some differences in the way they display facial expressions as well. This result is consistent to the findings of a previous study for children with ASD (Del Coco et al., 2017). For ADHD, previous studies have observed that people with this condition may find it difficult to recognize emotions from facial expressions (Da Fonseca et al., 2009). Hence, it is a possibility that people with ADHD may themselves display expressions in a different way from Control group, which is confirmed by this study.

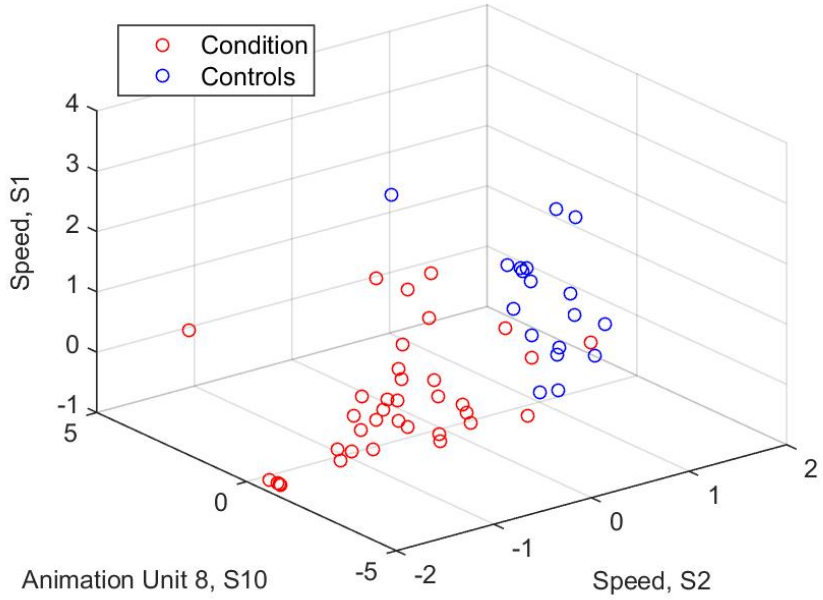


Figure 6.4: Top 3 features distinguishing Condition (ASD/ADHD) from control group. Animation Unit 8 corresponds to lip-corner depressor. S1, S2, S10 denote video segments corresponding to story 1, 2 and 10 of the 'Strange Stories' task respectively.

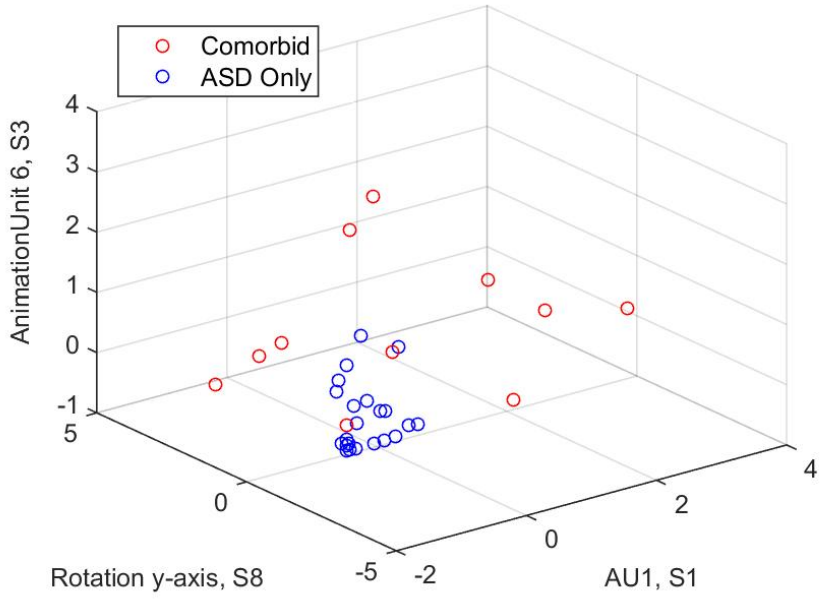


Figure 6.5: Top 3 features distinguishing Comorbid (ASD+ADHD) from ASD only group. Animation Unit 6 and AU1 corresponds to lip-corner puller and inner-brow raiser respectively. S1, S3 and S8 denote video segments corresponding to story 1, 3 and 8 of the 'Strange Stories' task respectively.

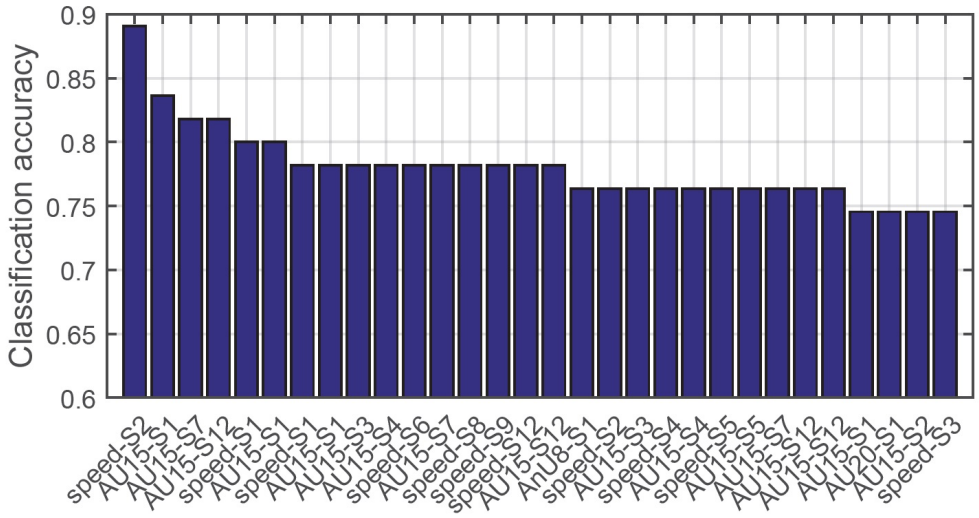


Figure 6.6: Top 30 features for classification of Controls vs Condition group. Each feature is represented by its feature type followed by the video segment number it was computed on. For e.g. AU15-S1 means that the feature corresponds to AU15 intensity histogram computed from the video segment corresponding to story 1 of the 'Strange stories' task. Please note that the same feature name can appear more than once because they are different features coming from the same histogram.

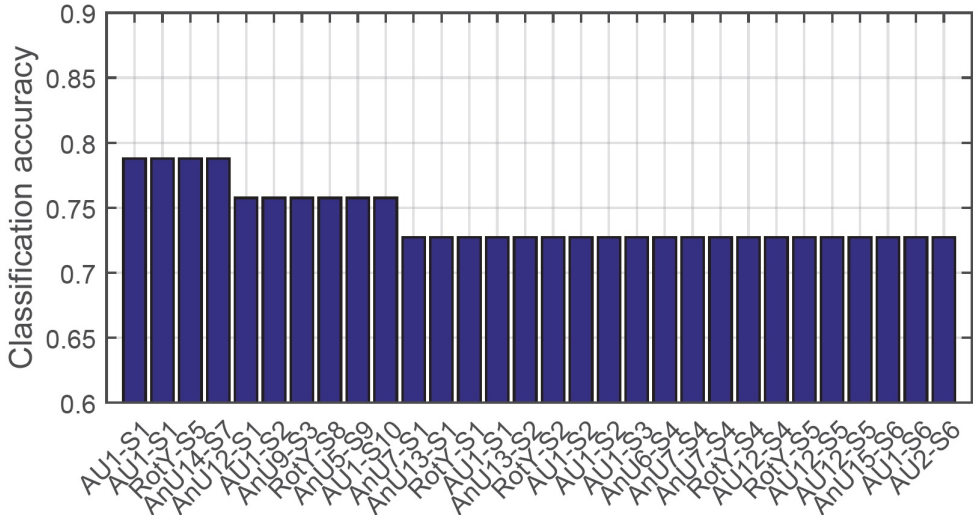


Figure 6.7: Top 30 features for classification of ASD vs Comorbid group. Features are named in the same way as in Fig. 6.6.

### 6.2.2 Distribution of features among different groups

In order to visualize the distribution of important sets of features in different groups of subjects, the average values of these features is plotted for each group. These features were extracted by computing histograms over entire (unsegmented) video of each subject, followed by Z-score normalization. For visualization purposes, an offset of 1.0 (for AnU8 and Speed) and 0.5 (for Rotation about Y-axis and AU1 intensity) was added to these features to make them all positive. Fig. 6.8 (first row) shows the distributions of head motion speed and AnU8 (lip corner depressor), respectively, in the control and condition group. For the head motion speed, it can observed that people in the condition group move their head at either higher speeds or at very low speeds (close to zero) much more as compared to people in the control group who displayed most of their head motion at medium speeds (in the range of 0.1-0.3 cm/sec). For AnU8 (lip-corner depressor) intensities, it can be observed that people in the condition group show higher intensities of this Animation Unit many more times than people in the control group. The distributions of head motion speed and AnU8 intensities show that these features provide a high discriminative power for classification of healthy controls and people with condition (ADHD/ASD).

Fig. 6.8 (second row) shows the distributions of head rotation (about vertical axis) and AU1 intensities respectively, in the ASD and comorbid group. In case of head rotation, it can be observed that people in the comorbid group deviate away from the screen to large extent many more times compared to people in the ASD group. This result is expected of the people in the comorbid group because of the presence of ADHD condition. People with ADHD can easily lose attention or move their head due to their hyperactive nature (2 core symptoms of ADHD), resulting in the deviation of head away from the screen. In case of AU1 intensities, it can be observed that people in the comorbid group also show higher intensities of AU1 many more times as compared to people in the ASD group. These observations show that both head rotation and AU1 intensities provide a high discriminative power for classification of ASD and comorbid (ADHD+ASD) conditions.

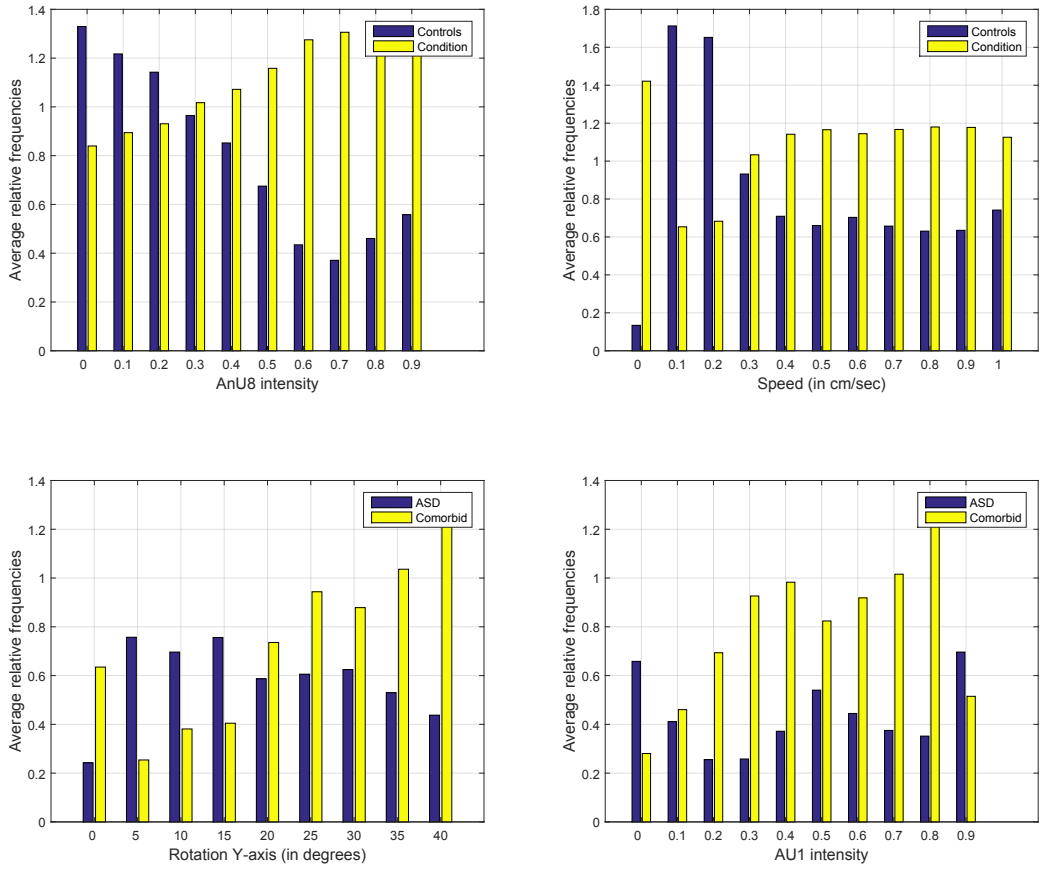


Figure 6.8: Visualization of average histogram (Z-score normalized) of the most discriminative features found for each classification problem. These histograms were computed over all video segments (entire video). The first row shows the histograms of head speed and AnU8 intensities (lip corner depressor), which were found to be discriminative for controls vs condition classification. The second row shows the histograms for head-rotation (about Y axis) and AU1 intensities, which were found to be most discriminative for ASD vs Comorbid group classification.

### 6.2.3 Effect of video segmentation

In order to evaluate the effect of splitting each subject's video into smaller segments and then extracting features from them separately as opposed to extracting them over the entire video, a set of baseline models were trained in which the features were extracted without splitting them into smaller segments. Fig. 6.9 shows a comparison of the performance of these models. In this Fig. it can be clearly seen that the proposed approach of splitting each video into separate segments each corresponding to a story in the Strange Stories task, which is aimed to contextualize the facial behaviour, results in a higher performance as compared to the baseline models. The performance was significantly higher, in particular for ASD vs Comorbid classification. For controls vs condition classification, only a marginal improvement was observed. This could be due to the fact that the most discriminative feature found for controls vs condition group classification, was the speed of head motion which might not be much affected by the context of the ongoing task. On the other hand, the discriminative features found for ASD vs Comorbid group classification includes AU1 intensity and head rotation (about Y axis) which may be much more dependent on the context of the task.

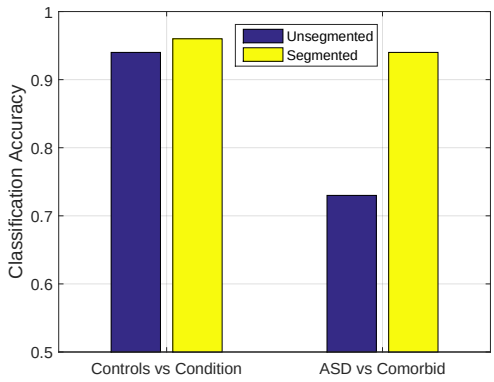


Figure 6.9: Performance comparison between segmented and unsegmented (no video splitting) methods of feature extraction.

### 6.2.4 Predictive power of individual video segments

An experiment was conducted in which the features from the individual video segments corresponding to each story in the Strange stories task were used to train separate models for both classification problems (controls vs condition and ASD vs comorbid). This was

done to test the facial behaviour response from which stories are more discriminative for classification. Similar to the previous experiments, in this set of experiments also the feature selection was done on the entire dataset and the performance of the classifiers were evaluated using leave-one-subject-out cross validation. Fig. 6.10 and 6.11 show the classification accuracies obtained for the 2 classification problems, when each video segment is used independently. For classification of control and condition group, segments corresponding to story 2 and 4 can be observed to be most discriminative, while for classification of ASD and comorbid group, segments corresponding to story 7 and 11 can be observed to be most discriminative. These results show that the context of a specific task (represented by each video segment) can play an important role in discriminating each group of participants. Since each video segment here consists of a participant reading and responding to a specific story (from the Strange Stories Test), the results indicate that some of these stories may cause a facial behaviour response in participants which can be more discriminative for their classification. It is also interesting to note that the story 4 (which consists of a joke) and story 7 (which consists of a double bluff) have been reported to be relatively more difficult (in comparison to other stories) to be correctly answered by people with ASD (Jolliffe and Baron-Cohen, 1999). This relative difficulty in understanding these stories could be causing an emotional response and result in discriminative facial behaviour. This hypothesis could be one of the reasons why the video segments of certain stories have been found to have more classification power than others.

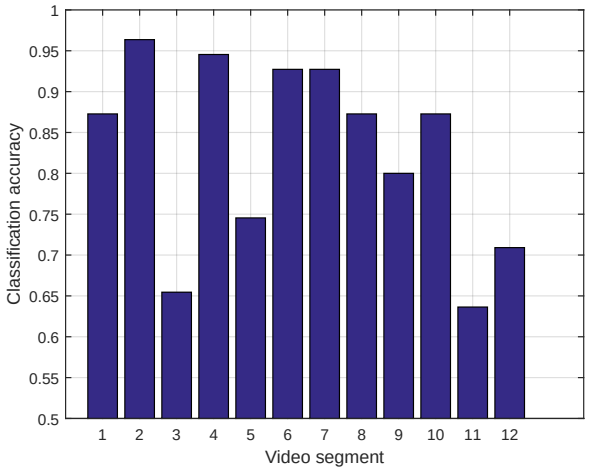


Figure 6.10: Predictive power of individual video segments (corresponding to each story in the Strange stories task) for classification of control and condition group.

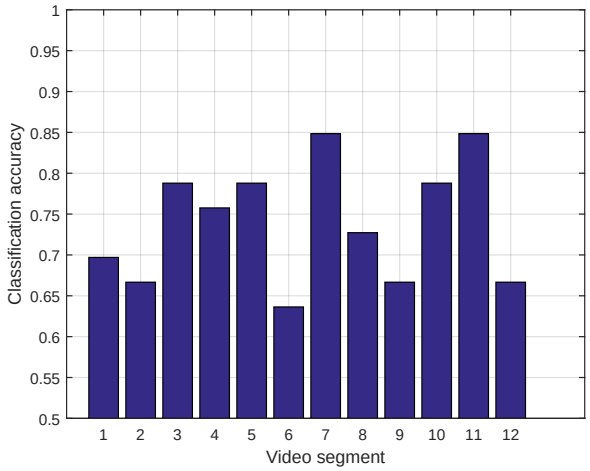


Figure 6.11: Predictive power of individual video segments (corresponding to each story in the Strange stories task) for classification of ASD and comorbid group.

## 6.3 Conclusion

This chapter described a novel method for making diagnostic prediction of ASD and comorbid (ADHD+ASD) conditions using automatic analysis of facial behaviour. Facial cues such as head motion, facial expression and pose are used in learning models which can accurately predict ADHD and ASD. The potential of facial expressions and other facial behavioural features was investigated for classification of individuals with these conditions. A high performance was achieved for classification of condition and healthy controls and for ASD and comorbid (ADHD+ASD) conditions. Due to the small size of the dataset, the dimensionality reduction step had to be done on the entire dataset (including test data). Doing dimensionality reduction using all subjects is not very uncommon and have been previously done (e.g. Mavadati et al. (2013)) especially when small number of subjects are involved. Due to this, the classification accuracies achieved may be towards the optimistic side as compared to the true accuracy achievable with this approach, which can only be estimated on a very large dataset. However, the main contribution of this approach is to show that the proposed facial behavioural features provide a high discriminative power which can help in the prediction of these conditions. These features are able to capture the subtle differences in the motion of head, pose of the head and facial expressions between people from the healthy control, ASD and the comorbid group, which consequently helps in distinguishing between these conditions. The context of facial behavioural response was also taken into account by splitting the videos

into smaller segments each corresponding to a story in the Strange Stories test. This enabled extraction of facial behavioural features in context to each specific story and was shown to improve the classification performance significantly. Overall, this approach shows promising potential and the initial results provide a proof of concept for its use as a diagnostic tool for ADHD, ASD and other neurodevelopmental conditions.

# Chapter 7

# Conclusions

This thesis proposed a dynamic deep learning framework for facial expression recognition and applied the same for facial behaviour analysis to aid diagnosis of ADHD and ASD. This chapter summarises the main approaches proposed and the related experiments conducted in this thesis. It also discusses the limitations of the proposed approaches and gives directions for future work.

## 7.1 Facial expression recognition

For facial expression recognition, the main focus was to detect the occurrence and intensity of facial Action Units (AU) automatically for each frame in a given video. In Chapter 4, the proposed deep learning framework was aimed at incorporating the three important characteristics which distinguishes one AU from another: shape, appearance and dynamics. Shape was encoded through binary masks computed from facial landmarks, appearance was encoded by the image region relevant for each AU. The dynamics was encoded by using a short input image sequence to CNN (for learning short term dynamics) and using BLSTM (for learning long term dynamics). The approach was evaluated on a number of databases (SEMAINE, BP4D and DISFA) and was found to achieve state-of-art-performance for both occurrence and intensity estimation. The contribution of different components of the system such as the use of image regions, binary masks, images sequence as input to CNN and the use of BLSTM was measured. Each component was observed to contribute significantly in improving the overall performance.

Apart from the evaluation of different system components, the sensitivity of the system to variation in the frame-rate of test sequences, architectural parameters (number of pooling layers, dropout probability and FC layer size) and the input sequence length in CNN and BLSTM, were also analysed. It was observed that the performance of AU models was robust to changes in the test frame-rate (upto a factor of 0.5 times the frame-rate during training stage). However, decreasing the frame rate beyond a factor of 0.5, results in a significant drop in performance for some AUs. For architectural parameters, having additional pooling layers resulted in a drop in average performance with the performance dropping significantly when 2 additional pooling layers are used. Changing the dropout probability and FC (fully-connected) layer size, did not make any significant impact on the overall performance. On the other hand, the input sequence size in CNN and in BLSTM were shown to have significant impact for some AUs. It was also observed that training each AU using a separate network, although less efficient, gives better performance as compared to training them together in a single network. This points in the direction that at the level of shape, appearance and especially dynamics, each AU can be quite different from one another and consequently an approach in which a common set of features are used to model multiple AUs, may not work the best (at least for the current type of architecture).

**Limitations:**

One of the limitations of the proposed dynamic deep learning framework, is that the facial regions used for each AU, were defined manually using domain knowledge. This leaves room for subjective judgement and may not give the most optimum results. It was also observed that the learnt networks work well for frontal or near frontal faces. However, the accuracy can drop significantly if there is a large out of plane rotation of the face. This could be due to the fact that the databases used for training the networks contained very few images of faces with large out of plane rotation. Bigger databases containing faces with large head pose variation, could help in training models robust to pose variation. Occlusion due to presence of hair and eye-glasses were also found to be limiting the performance of the AU models. Another drawback of the proposed approach is that to achieve the best average performance, a separate network needs to be trained for each AU. This significantly increases the computation cost both at the training and test stage. It was also observed that the performance of the AU models can drop significantly if there is a large difference (more than a factor of 0.5) between the frame-rates of the training and test sequences. This is because the temporal information (e.g. rate of change

of appearance due to facial muscle actions) learnt during the training stage doesn't match with the temporal information encountered during the test stage.

**Future work:**
As a future work, it would be interesting to explore the development of an architecture which could learn multiple AUs together without any drop in performance or perhaps increase the performance by learning the correlations between different AUs. In the current work, there was no particular emphasis on tackling the problem of head-pose. Large pose variation (especially out of plane rotation) can severely degrade the accuracy of the current AU models. Techniques for head pose frontalization (e.g. Hassner et al. (2015)), joint learning of pose and AUs (e.g. Batista et al. (2017)), learning pose specific models (e.g. He et al. (2017)) and creating new strategies for augmenting the existing training data with more variation in pose, could help in addressing this problem. Automatic learning of facial regions relevant to each AU is another direction which needs more attention. Developing data driven approaches to automatically learn relevant facial regions could help in achieving more optimum performance. Another interesting direction would be to explore new loss functions for AU intensity estimation which are invariant to data imbalance. Such a loss function would enable training using larger databases without the need to do any sub-sampling. It would also be interesting to explore ways to learn AU models which are more robust to differences in the frame-rate during the training and test stage. There are some existing work which have explored this particular problem. For e.g. Sariyanidi et al. (2017) proposed an unsupervised learning framework to express facial expression variation as a linear combination of localised basis functions where the coefficients of these basis functions represent intensities of facial muscle movement. This approach was found to be robust to frame-rate variation in training and test sequences. However, more research is required to develop similar solutions for end-to-end deep learning frameworks.

## 7.2 Automatic detection of ADHD and ASD

The aim of this thesis was to apply the proposed facial expression algorithm to aid the diagnosis of neurodevelopmental conditions like ADHD and ASD. For this purpose, Chapter 5 introduced the KOMAA dataset which could be used for evaluating approaches that target behaviour analysis for automatic detection of ADHD and ASD. This database

consisted of videos of participants from control, ASD and ADHD groups, who were recorded while doing the ”Strange stories” task in front of a computer screen. This controlled recording scenario was designed to easily detect behaviour signals which can help discriminate one group from another. Apart from the video recordings, the database also consisted of metadata in the form of age, gender, diagnostic status, scores from screening questionnaires like AQ10 and ASRS and the face and body tracking data from Kinect. It also contains scores from another neuropsychiatric test: ”Reading the mind in the eyes” (RME) test.

An initial analysis into the dataset revealed that the number of participants who have diagnosed with ADHD only was far less than the number of participants in the other groups (controls, ASD and comorbid). Also, there was a difference in the gender distribution between the control and clinical group of participants. Future work in this area would be to recruit more participants not only to increase the overall size of the database but also to balance the number of participants and gender distribution in each group. The data from the screening questionnaire scores and RME test also indicated that it is much harder to distinguish between ASD and comorbid (ASD+ADHD) conditions as compared to distinguishing between healthy controls and people with either of these conditions. This indicated that there is an overlap between the symptoms captured by these measures and motivated the need to find alternative measures.

Finally, in Chapter 6, the KOMAA database was used to evaluate a new approach which applies automatic facial behaviour analysis for prediction of ADHD and ASD. Facial behavioural features were encoded using facial AU intensities (detected using the method described in Chapter 4), along with 3D face tracking and facial Animation Units data obtained from Kinect. In addition, the context of the behavioural responses was also encoded by splitting each video into smaller video segments and computing histogram based features separately for each segment. SVM based classification models were learnt from the proposed features and evaluated on the task of classifying people with condition (ASD/ADHD) from controls and people with ASD from people with comorbid condition (ADHD+ASD). High classification accuracies were observed for both tasks. However, as discussed in Chapter 6, it is difficult to estimate the true performance of the algorithm on such a small dataset. In addition, the feature selection step which was done on the entire dataset could also make the performance appear more optimistic on a small sized dataset. However, the main contribution of this work are the facial behavioural

features which were shown to provide high discriminative power for classification of these conditions. The discriminative power of these features was analysed in terms of classification accuracies, for each feature individually as well as a group for each video segment (corresponding to each story in the Strange stories task). The results showed that these facial behavioural features provide good discriminative and predictive power and can potentially be very useful in the diagnosis of ADHD and ASD.

**Limitations:**

The recording scenario used for creating the KOMAA database was designed to simulate real person to person conversation while at the same time keeping the system free from the need for any other human intervention. However, on-screen display of predefined text and playing a pre-recorded voice which reads the text out, may not be best way to simulate such an interaction and can prevent more naturalistic display of facial behaviour. Another drawback of the proposed system is that any problem with the accuracy of face tracking or AU detection (for e.g. due to non-frontal head-pose, occlusion from eye-glasses) can negatively impact the prediction performance. The performance of the proposed system has been evaluated only on a small group of people where the distribution of the people was skewed in terms of the medical diagnosis (ADHD/ASD/Controls) and gender. This prevents obtaining a true estimate of the system performance, if used in clinical practice. Lastly, the proposed system has been designed to be used only in a very specific setting which consists of people doing the Strange Stories test in front of a computer screen. Therefore, the current system cannot be used to extract people's facial behaviour encountered in common day to day situations, and utilize it to make diagnostic predictions.

**Future work:**

The approach described in this thesis can be considered as a stepping stone in the direction of automatic behaviour analysis and demonstrates the feasibility of using it for diagnosis of mental health conditions like ADHD and ASD. Future work in this direction would involve evaluating this approach on very large databases not only for ADHD and ASD but also for other mental health conditions like depression, anxiety, Tourette's syndrome, etc. The present work did not make use of behaviour signals from hand and foot movements mainly because it is difficult to track them reliably using the current algorithms. Developing new algorithms which can do precise tracking of the entire body will enable a more complete behaviour analysis as compared to facial behaviour alone.

Another interesting direction would be to develop new methods for representing human behavioural features which can efficiently encode temporal context as well as context of external stimuli. In order to develop automatic diagnostic systems which can utilise more naturalistic facial behaviour while at the same time do not require any additional human intervention, it is important to develop methods to simulate real human-human interaction. Developing virtual humans which can look, sound and behave like humans, could be one of the ways to simulate human-human interaction. This will help in bringing out a more realistic behaviour response from the people as they would normally do in a general setting involving interaction with real people. This is important if the ultimate aim is to develop automatic behaviour analysis systems which can be deployed in public places or even in people's homes to detect/monitor the symptoms of these conditions.

# Bibliography

Role of dopamine transporter genotype and maternal prenatal smoking in childhood hyperactive-impulsive, inattentive, and oppositional behaviors. *The Journal of Pediatrics*, 143(1):104 – 110, 2003.

Lenard A Adler, Thomas Spencer, Thomas E Brown, James Holdnack, Keith Saylor, Kory Schuh, Paula T Trzepacz, David W Williams, and Douglas Kelsey. Once-daily atomoxetine for adult attention-deficit/hyperactivity disorder: a 6-month, double-blind trial. *Journal of clinical psychopharmacology*, 29(1):44–50, 2009.

Hesham A Alabbasi, P Moldoveanu, and Alin Moldoveanu. Real time facial emotion recognition using kinect v2 sensor. *IOSR J. Comput. Eng. Ver. II*, 17(3):2278–2661, 2015.

Greg Allen and Eric Courchesne. Attention function and dysfunction in autism. *Frontiers in bioscience: a journal and virtual library*, 6:D105–19, 2001.

Carrie Allison, Bonnie Auyeung, and Simon Baron-Cohen. Toward brief âĂIJred flagsâĂİ for autism screening: the short autism spectrum quotient and the short quantitative checklist in 1,000 cases and 3,000 controls. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(2):202–212, 2012.

Timur Almaev, Brais Martinez, and Michel Valstar. Learning to transfer: transferring latent task structures and its application to person-specific facial action unit detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3774–3782, 2015.

Timur R Almaev and Michel F Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 356–361. IEEE, 2013.

American Psychiatric Association APA. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition: DSM-IV-TR®*. American Psychiatric Association, 2000.

American Psychiatric Association APA. *Diagnostic and Statistical Manual of Mental Disorders: Dsm-5*. Amer Psychiatric Pub Incorporated, 2013.

Philip Asherson, Ron Akehurst, JJ Sandra Kooij, Michael Huss, Kathleen Beusterien, Rahul Sasané, Shadi Gholizadeh, and Paul Hodgkins. Under diagnosis of adult adhd cultural influences and societal burden. *Journal of Attention Disorders*, 16(5 suppl): 20S–38S, 2012.

A. Asthana, S. Cheng, S. Zafeiriou, and M. Pantic. Robust discriminative response map fitting with constrained local models. 2013.

Gillian Baird, Emily Simonoff, Andrew Pickles, Susie Chandler, Tom Loucas, David Meldrum, and Tony Charman. Prevalence of disorders of the autism spectrum in a population cohort of children in south thames: the special needs and autism project (snap). *The lancet*, 368(9531):210–215, 2006.

Vineeth Nallure Balasubramanian, Jieping Ye, and Sethuraman Panchanathan. Biased manifold embedding: A framework for person-independent head pose estimation. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

Tadas Baltrusaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.

Tobias Banaschewski, Veit Roessner, Ralf W Dittmann, P Janardhanan Santosh, and Aribert Rothenberger. Non–stimulant medications in the treatment of adhd. *European child & adolescent psychiatry*, 13(1):i102–i116, 2004.

Tobias Banaschewski, David Coghill, Paramala Santosh, Alessandro Zuddas, Philip Asherson, Jan Buitelaar, Marina Danckaerts, Manfred Döpfner, Stephen V Faraone, Aribert Rothenberger, et al. Long-acting medications for the hyperkinetic disorders. *European child & adolescent psychiatry*, 15(8):476–495, 2006.

Russell A Barkley. Behavioral inhibition, sustained attention, and executive functions: constructing a unifying theory of adhd. *Psychological bulletin*, 121(1):65, 1997.

Simon Baron-Cohen, Therese Jolliffe, Catherine Mortimore, and Mary Robertson. Another advanced test of theory of mind: Evidence from very high functioning adults with autism or asperger syndrome. *Journal of Child psychology and Psychiatry*, 38 (7):813–822, 1997.

Simon Baron-Cohen, Sally Wheelwright, Jacqueline Hill, Yogini Raste, and Ian Plumb. The âĂIJreading the mind in the eyesâĂİ test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of child psychology and psychiatry*, 42(2):241–251, 2001a.

Simon Baron-Cohen, Sally Wheelwright, Richard Skinner, Joanne Martin, and Emma Clubley. The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians. *Journal of autism and developmental disorders*, 31(1):5–17, 2001b.

Júlio César Batista, Vítor Albiero, Olga RP Bellon, and Luciano Silva. Aumpnet: simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 866–871. IEEE, 2017.

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

David J Beymer. Face recognition under varying pose. Technical report, DTIC Document, 1993.

Joseph Biederman, Thomas Spencer, and Timothy Wilens. Evidence-based pharmacotherapy for attention-deficit hyperactivity disorder. *International Journal of Neuropsychopharmacology*, 7(1):77–97, 2004.

Christopher M Bishop. *Neural networks for pattern recognition*. Oxford university press, 1995.

Ran Breuer and Ron Kimmel. A deep learning perspective on the origin of facial expressions. *arXiv preprint arXiv:1705.01842*, 2017.

Thomas E Brown. *Brown Attention-deficit Disorder Scales: Adolescents and Adults*. Psychological Corporation, 1996.

Traolach S Brugha, Sally McManus, John Bankart, Fiona Scott, Susan Purdon, Jane Smith, Paul Bebbington, Rachel Jenkins, and Howard Meltzer. Epidemiology of autism spectrum disorders in adults in the community in england. *Archives of general psychiatry*, 68(5):459–465, 2011.

Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. pages 2887–2894, 2012.

Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. Learning partially-observed hidden conditional random fields for facial expression recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 533–540. IEEE, 2009.

Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European Conference on Computer Vision*, pages 109–122. Springer, 2014.

Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013a.

Wen-Sheng Chu, Fernando De la Torre, and Jeffery F Cohn. Selective transfer machine for personalized facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3515–3522, 2013b.

C Keith Conners, Drew Erhardt, and Elizabeth P Sparrow. Conners' adult adhd rating scales (caars): technical manual. MHS North Tonawanda, 1999.

T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. 2012.

T.F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision. Technical report, University of Manchester, 2004.

Ciprian Adrian Corneanu, Marc Oliu Simon, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

Rachel Craig, Elizabeth Fuller, Jennifer Mindell, and Sally Bridges. *Health Survey for England, 2014*. 2015.

David Da Fonseca, Valérie Seguier, Andreia Santos, François Poinso, and Christine Deruelle. Emotion understanding in children with adhd. *Child Psychiatry & Human Development*, 40(1):111–121, 2009.

Mohamed Dahmane and Jean Meunier. Emotion recognition using dynamic grid-based hog features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 884–888. IEEE, 2011.

M. Dantone, J. Gall, G. Fanelli, and L. J. Van Gool. Real-time facial feature detection using conditional regression forests. pages 2578–2585, 2012.

Charles Darwin, Paul Ekman, and Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

Dragos Datcu and Léon JM Rothkrantz. Facial expression recognition with relevance vector machines. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 193–196. IEEE, 2005.

Marco Del Coco, Marco Leo, Pierluigi Carcagnı, Paolo Spagnolo, Pier Luigi Mazzeo, Massimo Bernava, Flavia Marino, Giovanni Pioggia, and Cosimo Distante. A computer vision based approach for understanding emotional involvements in children with autism spectrum disorders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1401–1407, 2017.

Meltem Demirkus, Doina Precup, James J Clark, and Tal Arbel. Probabilistic temporal head pose estimation using a hierarchical graphical model. In *European Conference on Computer Vision*, pages 328–344. Springer, 2014.

Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883. IEEE, 2011.

P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. 2010.

Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE transactions on pattern analysis and machine intelligence*, 21(10):974–989, 1999.

Shaun M Eack, Deborah P Greenwald, Susan S Hogarty, Amber L Bahorik, Maralee Y Litschge, Carla A Mazefsky, and Nancy J Minshew. Cognitive enhancement therapy for adults with autism spectrum disorder: results of an 18-month feasibility study. *Journal of autism and developmental disorders*, 43(12):2866–2877, 2013.

P Ekman. Micro expression training tool (mett) online, 2004.

P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City (USA), 2002.

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.

Mostafa K Abd El Meguid and Martin D Levine. Fully automated recognition of spontaneous facial expressions in videos using random forest classifiers. *IEEE Transactions on Affective Computing*, 5(2):141–154, 2014.

Gianluca Esposito, Paola Venuti, Fabio Apicella, and Filippo Muratori. Analysis of unsupported gait in toddlers with autism. *Brain and Development*, 33(5):367 – 373, 2011.

Gabriele Fanelli, Juergen Gall, and Luc Van Gool. Real time head pose estimation with random regression forests. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 617–624. IEEE, 2011.

Hadeel Faras, Nahed Al Ateeqi, and Lee Tidmarsh. Autism spectrum disorders. *Annals of Saudi medicine*, 30(4):295, 2010.

Sachin Sudhakar Farfade, Mohammad J Saberian, and Li-Jia Li. Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 643–650. ACM, 2015.

B. Fasel. Head-pose invariant facial expression recognition using convolutional neural networks. In *Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on*, pages 529–534, 2002. doi: 10.1109/ICMI.2002.1167051.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

Fabian Flohr, Madalin Dumitru-Guzu, Julian FP Kooij, and Dariu M Gavrila. A probabilistic framework for joint pedestrian head and body orientation estimation. *IEEE Transactions on Intelligent Transportation Systems*, 16(4):1872–1882, 2015.

E Fonbonne. Epidemiology of autistic disorder and other pervasive developmental disorders. *Journal of Clinical Psychiatry*, 66:3–8, 2005.

Christine M Freitag. The genetics of autistic disorders and its clinical relevance: a review of the literature, 2007.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

Andrew Gee and Roberto Cipolla. Determining the gaze of faces in images. *Image and Vision Computing*, 12(10):639–647, 1994.

Julia Geissler and Klaus-Peter Lesch. A lifetime of attention-deficit/hyperactivity disorder: diagnostic challenges, treatment and neurobiological mechanisms. *Expert review of neurotherapeutics*, 11(10):1467–1484, 2011.

S. Ghosh, E. Laksana, S. Scherer, and L. P. Morency. A multi-label convolutional neural network approach to cross-domain action unit detection. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 609–615, Sept 2015.

J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D.P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *Proc. of FG*, Shanghai, China, April 2013. IEEE.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014. doi: 10.1109/CVPR.2014.81. URL `http://dx.doi.org/10.1109/CVPR.2014.81`.

Michael Glodek, Stephan Tschechne, Georg Layher, Martin Schels, Tobias Brosch, Stefan Scherer, Markus Kächele, Miriam Schmidt, Heiko Neumann, Günther Palm, et al. Multiple classifier systems for the classification of audio-visual emotional states. In *Affective Computing and Intelligent Interaction*, pages 359–368. Springer, 2011.

Nicolas Gourier, Daniela Hall, and James L Crowley. Estimating face orientation from robust detection of salient facial structures. In *FG Net Workshop on Visual Observation of Deictic Gestures*, volume 6, 2004.

Alex Graves and JÃijrgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, pages 5–6, 2005.

Amogh Gudi, H. Emrah Tasli, Tim M. den Uyl, and Andreas Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.

Liangke Gui, Tadas Baltrušaitis, and Louis-Philippe Morency. Curriculum learning for facial expression recognition. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 505–511. IEEE, 2017.

Joachim Hallmayer, Sue Cleveland, Andrea Torres, Jennifer Phillips, Brianne Cohen, Tiffany Torigoe, Janet Miller, Angie Fedele, Jack Collins, Karen Smith, et al. Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of general psychiatry*, 68(11):1095–1102, 2011.

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.

Shizhong Han, Zibo Meng, AHMED-SHEHAB KHAN, and Yan Tong. Incremental boosting convolutional neural network for facial action unit recognition. In *Advances in Neural Information Processing Systems*, pages 109–117, 2016.

Francesca GE Happé. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of autism and Developmental disorders*, 24(2):129–154, 1994.

Jordan Hashemi, Thiago Vallin Spina, Mariano Tepper, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, and Guillermo Sapiro. Computer vision tools for the non-invasive assessment of autism-related behavioral markers. *CoRR*, abs/1210.7014, 2012.

Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015.

Jun He, Dongliang Li, Bin Yang, Siming Cao, Bo Sun, and Lejun Yu. Multi view facial action unit detection based on cnn and blstm-rnn. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 848–853. IEEE, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Antonio HernÃąndez-Vela, Miguel Reyes, Laura Igual, Josep Moya, VerÃşnica Violant, and Sergio Escalera. Adhd indicators modelling based on dynamic time warping from rgbd data: a feasibility study. In *VI CVC Workshop on the progress of Research and Development, Barcelona, Computer Vision Center*, pages 59–62, 2011.

Nikolas Hesse, Tobias Gehrig, Hua Gao, and Hazım Kemal Ekenel. Multi-view facial expression recognition using local appearance features. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3533–3536. IEEE, 2012.

Eva Hesselmark, Stephanie Plenty, and Susanne Bejerot. Group cognitive behavioural therapy and group recreational activity for adults with autism spectrum disorders: A preliminary randomized controlled trial. *Autism*, 18(6):672–683, 2014.

S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9 (8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`.

Paul Hodgkins, L Eugene Arnold, Monica Shaw, Hervé Caci, Jennifer Kahle, Alisa G Woods, Susan Young, et al. A systematic review of global publication trends regarding long-term outcomes of adhd. *Frontiers in psychiatry*, 2:84, 2012.

Thanarat Horprasert, Yaser Yacoob, and Larry S Davis. Computing 3-d head orientation from a monocular image sequence. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 242–247. IEEE, 1996.

Jeffrey Huang, Xuhui Shao, and Harry Wechsler. Face pose discrimination using support vector machines (svm). In *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, volume 1, pages 154–156. IEEE, 1998.

Shashank Jaiswal, Brais Martinez, and Michel F Valstar. Learning to combine local models for facial action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–6. IEEE, 2015.

B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic. Decision level fusion of domain specific regions for facial action recognition. In *International Conference on Pattern Recognition*, 2014a.

Bihan Jiang, M.F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321, March 2011.

Bihan Jiang, Michel Valstar, Brais Martinez, and Maja Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE transactions on cybernetics*, 44(2):161–174, 2014b.

Therese Jolliffe and Simon Baron-Cohen. The strange stories test: A replication with high-functioning adults with autism or asperger syndrome. *Journal of autism and developmental disorders*, 29(5):395–406, 1999.

Michael Jones and Paul Viola. Fast multi-view face detection. *Mitsubishi Electric Research Lab TR-20003-96*, 3:14, 2003.

Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. An empirical exploration of recurrent network architectures. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2342–2350, 2015.

Heechul Jung, Sihaeng Lee, Sunjeong Park, Injae Lee, Chunghyun Ahn, and Junmo Kim. Deep temporal appearance-geometry network for facial expression recognition. *arXiv preprint arXiv:1503.01532*, 2015.

Brookes K, Mill J, Guindalini C, and et al. A common haplotype of the dopamine transporter gene associated with attention-deficit/hyperactivity disorder and interacting with maternal use of alcohol during pregnancy. *Archives of General Psychiatry*, 63(1):74–81, 2006.

Samira Ebrahimi Kahou, Christopher Pal, Xavier Bouthillier, Pierre Froumenty, Çaglar Gülçehre, Roland Memisevic, Pascal Vincent, Aaron Courville, Yoshua Bengio, Raul Chandias Ferrari, Mehdi Mirza, Sébastien Jean, Pierre-Luc Carrier, Yann Dauphin, Nicolas Boulanger-Lewandowski, Abhishek Aggarwal, Jeremie Zumer, Pascal Lamblin, Jean-Philippe Raymond, Guillaume Desjardins, Razvan Pascanu, David Warde-Farley, Atousa Torabi, Arjun Sharma, Emmanuel Bengio, Myriam Côté, Kishore Reddy Konda, and Zhenzhou Wu. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 543–550. ACM, 2013.

S. Kaltwang, S. Todorovic, and M. Pantic. Latent trees for estimating intensity of facial action units. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 296–304, June 2015.

Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing*, pages 368–377. Springer, 2012.

Hyunduk Kim, Sang-Heon Lee, Myoung-Kyu Sohn, and Dong-Ju Kim. Illumination invariant head pose estimation using random forests classifier and binary pattern run length matrix. *Human-Centric Computing and Information Sciences*, 4(1):9, 2014.

Minyoung Kim and Vladimir Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. In *European Conference on Computer Vision*, pages 649–662. Springer, 2010.

Sandra JJ Kooij, Susanne Bejerot, Andrew Blackwell, Herve Caci, Miquel Casas-Brugué, Pieter J Carpentier, Dan Edvinsson, John Fayyad, Karin Foeken, Michael Fitzgerald, et al. European consensus statement on diagnosis and treatment of adult adhd: The european network adult adhd. *BMC psychiatry*, 10(1):67, 2010.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. 2012a.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012b.

Jan-Olov Larsson, Henrik Larsson, and Paul Lichtenstein. Genetic and environmental contributions to stability and change of adhd symptoms between 8 and 13 years of age: a longitudinal twin study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43(10):1267–1275, 2004.

Yael Leitner. The co-occurrence of autism and attention deficit hyperactivity disorder in children–what do we know? *Brain Development and the Attention Spectrum*, page 80, 2007.

Haoxiang Li, Zhe Lin, Xiaohui Shen, Jonathan Brandt, and Gang Hua. A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5325–5334, 2015.

Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

Xiaobai Li, HONG Xiaopeng, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, and Matti Pietikainen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing*, 2017.

Yongmin Li, Shaogang Gong, and Heather Liddell. Support vector regression and classification based multi-view face detection and recognition. In *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on*, pages 300–305. IEEE, 2000.

Yongmin Li, Shaogang Gong, Jamie Sherrah, and Heather Liddell. Support vector machine based multi-view face detection and recognition. *Image and Vision Computing*, 22(5):413–427, 2004.

Jenn-Jier James Lien. *Automatic recognition of facial expressions using hidden markov models and estimation of exprssion intensity*. PhD thesis, Washington University, St. Louis, 1998.

Gwen Littlewort, Jacob Whitehill, Ting-Fan Wu, Nicholas Butko, Paul Ruvolo, Javier Movellan, and Marian Bartlett. The motion in emotionâĂŤa cert based approach to the fera emotion challenge. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 897–902. IEEE, 2011.

Mengyi Liu, Shaoxin Li, Shiguang Shan, Ruiping Wang, and Xilin Chen. Deeply learning deformable facial action parts model for dynamic expression analysis. In *ACCV 2014*, volume 9006 of *Lecture Notes in Computer Science*, pages 143–157. Springer International Publishing, 2015.

Fei Long, Tingfan Wu, Javier R Movellan, Marian S Bartlett, and Gwen Littlewort. Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. *Neurocomputing*, 93:126–132, 2012.

Catherine Lord, Michael Rutter, Susan Goode, Jacquelyn Heemsbergen, Heather Jordan, Lynn Mawhood, and Eric Schopler. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, 19(2):185–212, 1989.

Catherine Lord, Michael Rutter, and Ann Le Couteur. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, 24(5):659–685, 1994.

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified

expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *Trans. Sys., Man and Cybernetics, Part B*, 41:664–674, 2011.

Simon Lucey, Ahmed Bilal Ashraf, and Jeffrey F Cohn. *Investigating spontaneous facial action recognition through aam representations of the face*. INTECH Open Access Publisher, 2007.

Michael J Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, 1999.

Qi-rong Mao, Xin-yu Pan, Yong-zhao Zhan, and Xiang-jun Shen. Usingkinect for real-time emotion recognition via facial expressions. *Frontiers of Information Technology & Electronic Engineering*, 16(4):272–282, 2015.

B. Martinez, M. F. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression based facial point detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 35(5), pages 1149–1163, 2013.

Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014.

David Matsumoto and Hyi Sung Hwang. Evidence for training the ability to read microexpressions of emotion. *Motivation and Emotion*, 35(2):181–191, 2011.

Iain Matthews and Simon Baker. Active appearance models revisited. *Int. J. Comput. Vision*, 60(2):135–164, November 2004.

S Mohammad Mavadati and Mohammad H Mahoor. Temporal facial expression modeling for automated action unit intensity measurement. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4648–4653. IEEE, 2014.

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

Stephen J McKenna and Shaogang Gong. Real-time face pose estimation. *Real-Time Imaging*, 4(5):333–347, 1998.

G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroeder. The semaine database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3: 5–17, April 2012.

Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*, pages 69–82. Springer, 2004.

S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. pages 504–513, 2008.

Paulo A. V. Miranda, Alexandre X. Falcï£·o, and Jayaram K. Udupa. Cloud models: Their construction and employment in automatic mri segmentation of the brain.

S Moore and R Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541–558, 2011.

Sylvie Mrug, Brooke SG Molina, Betsy Hoza, Alyson C Gerdes, Stephen P Hinshaw, Lily Hechtman, and L Eugene Arnold. Peer rejection and friendships in children with attention-deficit/hyperactivity disorder: contributions to long-term outcomes. *Journal of abnormal child psychology*, 40(6):1013–1026, 2012.

Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2009.

Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. Output-associative rvm regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30 (3):186–196, 2012.

J. Nicolle, K. Bailly, and M. Chetouani. Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 06, pages 1–6, May 2015.

Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 501–508. ACM, 2012.

Joel Nigg, Molly Nikolas, and S Alexandra Burt. Measured gene-by-environment interaction in relation to attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(9):863–873, 2010.

Sourabh Niyogi and William T Freeman. Example-based head tracking. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 374–378. IEEE, 1996.

Timo Ojala, Matti Pietikainen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision & Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585. IEEE, 1994.

Curtis Padgett and Garrison W Cottrell. Representing face images for emotion classification. *Advances in neural information processing systems*, pages 894–900, 1997.

Igor S. Pandzic and Robert Forchheimer, editors. *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. John Wiley & Sons, Inc., New York, NY, USA, 2003. ISBN 0470854626.

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

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

Marc Parizeau and Rejean Plamondon. A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):710–717, 1990.

Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, volume 1, page 6, 2015.

Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1449–1456. IEEE, 2011.

Guilherme Polanczyk, Maurício Silva de Lima, Bernardo Lessa Horta, Joseph Biederman, and Luis Augusto Rohde. The worldwide prevalence of adhd: a systematic review and metaregression analysis. *American journal of psychiatry*, 2007.

Senya Polikovsky, Yoshinari Kameda, and Yuichi Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. 2009.

Shyam Sundar Rajagopalan and Roland Goecke. Detecting self-stimulatory behaviours for autism diagnosis. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 1470–1474. IEEE, 2014.

Shyam Sundar Rajagopalan, OV Ramana Murthy, Roland Goecke, and Agata Rozga. Play with meâĂŤmeasuring a child's engagement in a social interaction. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. A deep pyramid deformable part model for face detection. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8. IEEE, 2015.

Patricia A Rao and Rebecca J Landa. Association between severity of behavioral phenotype and comorbid attention deficit hyperactivity disorder symptoms in children with autism spectrum disorders. *Autism*, page 1362361312470494, 2013.

Patricia A Rao and Rebecca J Landa. Association between severity of behavioral phenotype and comorbid attention deficit hyperactivity disorder symptoms in children with autism spectrum disorders. *Autism*, 18(3):272–280, 2014.

Bisser Raytchev, Ikushi Yoda, and Katsuhiko Sakaue. Head pose estimation by nonlinear manifold learning. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 4, pages 462–466. IEEE, 2004.

James M Rehg. Behavior imaging: Using computer vision to study autism. *MVA*, 11: 14–21, 2011.

Fuji Ren and Zhong Huang. Facial expression recognition based on aam–sift and adaptive regional weighting. *IEEJ Transactions on Electrical and Electronic Engineering*, 10(6):713–722, 2015.

JA Ressel. A circumplex model of affect. *J. Personality and Social Psychology*, 39: 1161–78, 1980.

Salah Rifai, Yoshua Bengio, Aaron Courville, Pascal Vincent, and Mehdi Mirza. Disentangling factors of variation for facial expression recognition. In *ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pages 808–822. Springer Berlin Heidelberg, 2012.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Multi-output laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2634–2641. IEEE, 2012.

Ailsa J Russell, Amita Jassi, Miguel A Fullana, Hilary Mack, Kate Johnston, Isobel Heyman, Declan G Murphy, and David Mataix-Cols. Cognitive behavior therapy for comorbid obsessive-compulsive disorder in high-functioning autism spectrum disorders: A randomized controlled trial. *Depression and anxiety*, 30(8):697–708, 2013.

Enrique Sánchez-Lozano, Brais Martinez, Georgios Tzimiropoulos, and Michel Valstar. Cascaded continuous regression for real-time incremental face tracking. In *European Conference on Computer Vision*, pages 645–661. Springer, 2016.

Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image and Vision Computing*, 30(10):683 – 697, 2012.

Jason M. Saragih, Simon Lucey, and Jeffrey F. Cohn. Face alignment through subspace constrained mean-shifts. In *proc. ACMInt'l Conf. ICCV*, pages 1034–1041. IEEE, 2009.

Evangelos Sariyanidi, Hatice Gunes, Muhittin Gökmen, and Andrea Cavallaro. Local zernike moment representation for facial affect recognition. In *BMVC*, 2013.

Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1113–1133, 2015.

Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Learning bases of activity for facial expression recognition. *IEEE Transactions on Image Processing*, 26(4): 1965–1978, 2017.

Amy M Schatz, Amy K Weimer, and Doris A Trauner. Brief report: Attention differences in asperger syndrome. *Journal of autism and developmental disorders*, 32(4):333–336, 2002.

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

Henry Schneiderman and Takeo Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.

Edgar Seemann, Kai Nickel, and Rainer Stiefelhagen. Head pose estimation using stereo vision for human-robot interaction. In *Automatic Face and Gesture Recognition, 2004. Proceedings. Sixth IEEE International Conference on*, pages 626–631. IEEE, 2004.

Thibaud Senechal, Vincent Rapp, and Lionel Prevost. Facial feature tracking for emotional dynamic analysis. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 495–506. Springer, 2011.

Caifeng Shan. *Inferring facial and body language*. PhD thesis, Queen Mary University of London, 2008.

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.

Evan Shelhamer, Jonathon Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

Jamie Sherrah, Shaogang Gong, and Eng-Jon Ong. Understanding pose discrimination in similarity space. In *BMVC*, pages 1–10. Citeseer, 1999.

J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304, Washington, DC, USA, 2011.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86:420–428, 1979.

Viktória Simon, Pál Czobor, Sára Bálint, Ágnes Mészáros, and István Bitter. Prevalence and correlates of adult attention-deficit hyperactivity disorder: meta-analysis. *The British Journal of Psychiatry*, 194(3):204–211, 2009.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ravishankar Sivalingam, Anoop Cherian, Joshua Fasching, Nicholas Walczak, Nathaniel D. Bird, Vassilios Morellas, Barbara Murphy, Kathryn Cullen, Kelvin O. Lim, Guillermo Sapiro, and Nikolaos Papanikolopoulos. A multi-sensor visual tracking system for behavior monitoring of at-risk children. In *ICRA*, pages 1345–1350. IEEE, 2012.

H Soyel and H Demirel. Facial expression recognition based on discriminative scale invariant feature transform. *Electronics letters*, 46(5):343–345, 2010.

Debbie Spain, Jacqueline Sin, Trudie Chalder, Declan Murphy, and Francesca Happe. Cognitive behaviour therapy for adults with autism spectrum disorders and psychiatric co-morbidity: A review. *Research in Autism Spectrum Disorders*, 9:151–162, 2015.

Annelies A Spek, Nadia C Van Ham, and Ivan Nyklíček. Mindfulness-based therapy in adults with an autism spectrum disorder: a randomized controlled trial. *Research in developmental disabilities*, 34(1):246–253, 2013.

Thomas Spencer, Joseph Biederman, Timothy Wilens, Robert Doyle, Craig Surman, Jefferson Prince, Eric Mick, Megan Aleardi, Kathleen Herzig, and Stephen Faraone. A large, double-blind, randomized clinical trial of methylphenidate in the treatment of adults with attention-deficit/hyperactivity disorder. *Biological psychiatry*, 57(5): 456–463, 2005.

Thomas J Spencer, Joseph Biederman, and Eric Mick. Attention-deficit/hyperactivity disorder: diagnosis, lifespan, comorbidities, and neurobiology. *Journal of pediatric psychology*, 32(6):631–642, 2007.

Sujith Srinivasan and Kim L Boyer. Head pose estimation using view based eigenspaces. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 302–305. IEEE, 2002.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Jim Stevenson, Phil Asherson, David Hay, Florence Levy, Jim Swanson, Anita Thapar, and Erik Willcutt. Characterizing the adhd phenotype for genetic studies. *Developmental Science*, 8(2):115–121, 2005.

Rainer Stiefelhagen. Estimating head pose with neural networks-results on the pointing04 icpr workshop evaluation data. In *Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures*, volume 1, 2004.

Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

J Tarver, D Daley, and K Sayal. Attention-deficit hyperactivity disorder (adhd): an updated review of the essential facts. *Child: care, health and development*, 40(6):762–774, 2014.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Anita Thapar, Miriam Cooper, Olga Eyre, and Kate Langley. Practitioner review: what have we learnt about the causes of adhd? *Journal of Child Psychology and Psychiatry*, 54(1):3–16, 2013.

Yan Tong, Jixu Chen, and Qiang Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *IEEE transactions on pattern analysis and machine intelligence*, 32(2):258–273, 2010.

Zoltán Tősér, László A Jeni, András Lőrincz, and Jeffrey F Cohn. Deep learning for facial action unit detection under large head poses. In *Computer Vision–ECCV 2016 Workshops*, pages 359–371. Springer, 2016.

Christopher W Tyler and Chien-Chung Chen. Signal detection theory in the 2afc paradigm: attention, channel uncertainty and probability summation. *Vision research*, 40(22):3121–3144, 2000.

M. F. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. pages 2729–2736, 2010.

Michel Valstar. Automatic behaviour understanding in medicine. In *Proc. of RFMIR, ICMI*, pages 57–60, Istambul, Turkey, November 2014. ACM.

Michel Valstar, Jeff Girard, Timur Almaev, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeff Cohn. Fera 2015 - second facial expression recognition and analysis challenge. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015a.

Michel Valstar, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeff Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In *Proc. of FG*, Ljublijana, Slovenia, May 2015b. IEEE.

Michel F Valstar and Maja Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *International Workshop on Human-Computer Interaction*, pages 118–127. Springer, 2007.

Michel F Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1):28–43, 2012a.

Michel F Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(1):28–43, 2012b.

Michel F Valstar, Timur Almaev, Jeffrey M Girard, Gary McKeown, Marc Mehu, Lijun Yin, Maja Pantic, and Jeffrey F Cohn. Fera 2015-second facial expression recognition and analysis challenge. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–8. IEEE, 2015c.

Chelsea A Vaughan. Test review: E. schopler, me van bourgondien, gj wellman, & sr love childhood autism rating scale . los angeles, ca: Western psychological services, 2010. *Journal of Psychoeducational Assessment*, 29(5):489–493, 2011.

A. Vedaldi and K. Lenc. Matconvnet: Convolutional neural networks for matlab. In *ACM International Conference on Multimedia*, 2015.

Michael Voit, Kai Nickel, and Rainer Stiefelhagen. Neural network-based head pose estimation and multi-view fusion. In *International Evaluation Workshop on Classification of Events, Activities and Relationships*, pages 291–298. Springer, 2006.

Michael Voit, Kai Nickel, and Rainer Stiefelhagen. Head pose estimation in single- and multi-view environments-results on the clearâĂŹ07 benchmarks. In *Multimodal Technologies for Perception of Humans*, pages 307–316. Springer, 2008.

Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 831–834. ACM, 2013.

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.

Peng Wang, Frederick Barrett, Elizabeth Martin, Marina Milonova, Raquel E. Gur, Ruben C. Gur, Christian Kohler, and Ragini Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of Neuroscience Methods*, 168(1):224 – 238, 2008.

Ziheng Wang, Yongqiang Li, Shangfei Wang, and Qiang Ji. Capturing global semantic relationships for facial action unit recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3304–3311, 2013.

Mark F Ward. The wender utah rating scale: an aid in the retrospective. *Am j psychiatry*, 1(50):885, 1993.

Felix Weninger, Johannes Bergmann, and Björn Schuller. Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *J. Mach. Learn. Res.*, 16 (1):547–551, January 2015.

P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, Oct 1990.

PJ Werbos. Beyond regression: new tools for prediction and analysis in the behavioral sciences [ph. d. thesis] cambridge. *Mass, USA: Hardward University*, 1974.

Jacob Whitehill and Christian W Omlin. Haar features for facs au recognition. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 5–pp. IEEE, 2006.

Lorna Wing, Susan R Leekam, Sarah J Libby, Judith Gould, and Michael Larcombe. The diagnostic interview for social and communication disorders: Background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry*, 43(3): 307–325, 2002.

Qi Wu, Xunbing Shen, and Xiaolan Fu. The machine knows what you are hiding: an automatic micro-expression recognition system. *Affective Computing and Intelligent Interaction*, pages 152–162, 2011a.

T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S.Bartlett, and Javier R. Movellan. Multi-layer architectures of facial action unit recognition. 2012. In print.

Tingfan Wu, Marian S Bartlett, and Javier R Movellan. Facial expression recognition using gabor motion energy filters. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 42–47. IEEE, 2010.

Tingfan Wu, Nicholas J Butko, Paul Ruvolo, Jacob Whitehill, Marian S Bartlett, and Javier R Movellan. Action unit recognition transfer across datasets. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 889–896. IEEE, 2011b.

Xuehan Xiong and Fernando De la Torre. Global supervised descent method. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2664–2673, 2015.

Xuehan Xiong and Fernando De la Torre Frade. Supervised descent method and its applications to face alignment. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2013.

Junjie Yan, Zhen Lei, Dong Yi, and Stan Li. Learn to combine multiple hypotheses for accurate face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 392–396, 2013a.

Wen-Jing Yan, Qi Wu, Jing Liang, Yu-Hsin Chen, and Xiaolan Fu. How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, 37(4):217–230, 2013b.

Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9(1):e86041, 2014.

Yan Yan, Elisa Ricci, Ramanathan Subramanian, Gaowen Liu, Oswald Lanz, and Nicu Sebe. A multi-task learning framework for head pose estimation under target motion. *IEEE transactions on pattern analysis and machine intelligence*, 38(6):1070–1083, 2016.

Peng Yang, Qingshan Liu, and Dimitris N Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

Peng Yang, Qingshan Liu, and Dimitris Metaxas. Dynamic soft encoded patterns for facial event analysis. *Computer Vision and Image Understanding*, 115(3):456–465, 2011.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3676–3684, 2015.

Benjamin E Yerys, Gregory L Wallace, Jennifer L Sokoloff, Devon A Shook, Joette D James, and Lauren Kenworthy. Attention deficit/hyperactivity disorder symptoms

moderate cognition and behavior in children with autism spectrum disorders. *Autism Research*, 2(6):322–333, 2009.

Anil Yüce, Hua Gao, and Jean-Philippe Thiran. Discriminant multi-label manifold embedding for facial action unit detection. In *Facial Expression Recognition and Analysis Challenge, in conjunction with IEEE Int'l Conf. on Face and Gesture Recognition*, 2015.

Stefanos Zafeiriou, Cha Zhang, and Zhengyou Zhang. A survey on face detection in the wild: past, present and future. *Computer Vision and Image Understanding*, 138:1–24, 2015.

Stefanos Zafeiriou, Athanasios Papaioannou, Irene Kotsia, Mihalis Nicolaou, and Guoying Zhao. Facial affect"in-the-wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 36–47, 2016.

Cha Zhang and Zhengyou Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1036–1041. IEEE, 2014.

Wenchao Zhang, Shiguang Shan, Wen Gao, Xilin Chen, and Hongming Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. In *ICCV*, volume 1, pages 786–791 Vol. 1, 2005.

Xiao Zhang, Mohammad H Mahoor, S Mohammad Mavadati, and Jeffrey F Cohn. A l p-norm mtmkl framework for simultaneous detection of multiple facial action units. In *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, pages 1104–1111. IEEE, 2014a.

Xing Zhang, Lijun Yin, Jeffrey F. Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692 – 706, 2014b.

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6), 2007.

Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and Dimitris N Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2562–2569. IEEE, 2012.

Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4998–5006, 2015.

Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, pages 2879–2886, 2012a.

Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012b.

Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879–2886. IEEE, 2012c.

Yunfeng Zhu, F. De la Torre, J.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, April 2011.

# Appendix A

# Screening questionnaires for ADHD and ASD

NHS National Institute for Health Research

# AQ-10

## Autism Spectrum Quotient (AQ)

*A quick referral guide for adults with suspected autism who do not have a learning disability.*

| ***Please tick one option per question only:*** | | **Definitely Agree** | **Slightly Agree** | **Slightly Disagree** | **Definitely Disagree** |
|---|---|---|---|---|---|
| 1 | I often notice small sounds when others do not | | | | |
| 2 | I usually concentrate more on the whole picture, rather than the small details | | | | |
| 3 | I find it easy to do more than one thing at once | | | | |
| 4 | If there is an interruption, I can switch back to what I was doing very quickly | | | | |
| 5 | I find it easy to 'read between the lines' when someone is talking to me | | | | |
| 6 | I know how to tell if someone listening to me is getting bored | | | | |
| 7 | When I'm reading a story I find it difficult to work out the characters' intentions | | | | |
| 8 | I like to collect information about categories of things (e.g. types of car, types of bird, types of train, types of plant etc) | | | | |
| 9 | I find it easy to work out what someone is thinking or feeling just by looking at their face | | | | |
| 10 | I find it difficult to work out people's intentions | | | | |

**SCORING:** Only 1 point can be scored for each question. *Score 1 point for Definitely or Slightly Agree on each of items 1, 7, 8, and 10. Score 1 point for Definitely or Slightly Disagree on each of items 2, 3, 4, 5, 6, and 9.* If the individual scores **more than 6 out of 10**, consider referring them for a specialist diagnostic assessment.

This test is recommended in 'Autism: recognition, referral, diagnosis and management of adults on the autism spectrum' (NICE clinical guideline CG142). www.nice.org.uk/CG142

**Key reference:** Allison C, Auyeung B, and Baron-Cohen S, (2012) *Journal of the American Academy of Child and Adolescent Psychiatry* 51(2):202-12.

UNIVERSITY OF CAMBRIDGE

autism research centre arc

# Adult ADHD Self-Report Scale (ASRS-v1.1) Symptom Checklist

| Patient Name | | Today's Date | |
|---|---|---|---|

| Please answer the questions below, rating yourself on each of the criteria shown using the scale on the right side of the page. As you answer each question, place an X in the box that best describes how you have felt and conducted yourself over the past 6 months. Please give this completed checklist to your healthcare professional to discuss during today's appointment. | Never | Rarely | Sometimes | Often | Very Often |
|---|---|---|---|---|---|
| 1. How often do you have trouble wrapping up the final details of a project, once the challenging parts have been done? | | | | | |
| 2. How often do you have difficulty getting things in order when you have to do a task that requires organization? | | | | | |
| 3. How often do you have problems remembering appointments or obligations? | | | | | |
| 4. When you have a task that requires a lot of thought, how often do you avoid or delay getting started? | | | | | |
| 5. How often do you fidget or squirm with your hands or feet when you have to sit down for a long time? | | | | | |
| 6. How often do you feel overly active and compelled to do things, like you were driven by a motor? | | | | | |
| | | | | | Part A |
| 7. How often do you make careless mistakes when you have to work on a boring or difficult project? | | | | | |
| 8. How often do you have difficulty keeping your attention when you are doing boring or repetitive work? | | | | | |
| 9. How often do you have difficulty concentrating on what people say to you, even when they are speaking to you directly? | | | | | |
| 10. How often do you misplace or have difficulty finding things at home or at work? | | | | | |
| 11. How often are you distracted by activity or noise around you? | | | | | |
| 12. How often do you leave your seat in meetings or other situations in which you are expected to remain seated? | | | | | |
| 13. How often do you feel restless or fidgety? | | | | | |
| 14. How often do you have difficulty unwinding and relaxing when you have time to yourself? | | | | | |
| 15. How often do you find yourself talking too much when you are in social situations? | | | | | |
| 16. When you're in a conversation, how often do you find yourself finishing the sentences of the people you are talking to, before they can finish them themselves? | | | | | |
| 17. How often do you have difficulty waiting your turn in situations when turn taking is required? | | | | | |
| 18. How often do you interrupt others when they are busy? | | | | | |
| | | | | | Part B |

# Appendix B

# Strange Stories Test

One day, while she is playing in the house, Anna accidently knocks over and breaks her mother's favourite crystal vase. Oh dear, when mother finds out she'll be very cross. So when Anna's mother comes home and sees the broken vase and asks Anna what happened, Anna says "The dog knocked it over, it wasn't my fault!".

- Q1. Was it true, what Anna told her mother?
- Q2. Why did she say this?

One day Aunt Jane came to visit Peter. Now Peter loves his Aunt very much, but today she is wearing a new hat; a new hat which Peter thinks is very ugly indeed. Peter thinks his Aunt looks silly in it and much nicer in her old. But when Aunt Jane asks Peter, "How do you like my new hat?", Peter says, "Oh, its very nice".

Q1. Was it true what Peter said?

Q2. Why did he say it?

Katie and Emma are playing in the house Emma picks up a banana from the fruit bowl and hold it up to her ear. She says to Katie, "Look! This banana is a telephone!"

Q1. Is it true what Emma says?

Q2. Why does Emma say this?

Daniel and Ian see Mrs Thompson coming out of the hairdresser's one day. She looks a bit funny because the hairdresser has cut her hair much too short. Daniel says to Ian, "She must have been in a fight with a lawnmower!".

Q1. Is it true, what Daniel says?

Q2. Why does he say this?

Emma has a cough. All through lunch she coughs and coughs. Father says, "Poor Emma, you must have a frog in your throat!".

Q1. Is it true, what Father says to Emma?

Q2. Why does he say that?

A burglar who has just robbed a shop is making his getaway. As he is running home, a policeman on his beat sees him drop his glove. He doesn't know the man is a burglar, he just wants to tell him he dropped his glove. But when the policeman shouts out to the burglar, "Hey you! Stop!", the burglar turns round, sees the policeman and gives himself up. He puts his hand up and admits he did the break-in at the local shop.

Q1. Was the policeman surprised by what the burglar did?

Q2. Why did the burglar do this, when the policeman just wanted to give him back his glove?

During the war, the Red army capture a member of the Blue army. They want him to tell them where his army's tanks are; they know they are either by the sea or in the mountains. They know that the prisoner will not want to tell them, he will want to save army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his army's tanks. Now when the other side ask him where his tanks are, he says, "they are in the mountains".

Q1. Is it true, what the prisoner said?
Q2. Where will the other army look for his tanks?
Q3. Why did the prisoner say what he said?

Sarah, and Tom are going on a picnic. It is Tom's idea, he says it is going to be a lovely sunny day for a picnic. But just as they are unpacking the food, it starts to rain, and soon they are both soaked to the skin. Sarah is cross. She says, "Oh yes, a lovely day for a picnic alright!".

Q1. Is it true, what Sarah says?
Q2. Why does she say this?

Brian is always hungry. Today at school it is his favourite meal – sausages and beans. He is a very greedy boy, and he would like to have more sausages than anybody else, even though his mother will have made him a lovely tea when he gets home! But everyone is allowed two sausages and no more. When it is Brian's turn to be served, he says, "Oh please can I have 4 sausages, because I won't be having any tea when I get home!".

Q1. Is it true, what Brian says?
Q2. Why does he say that?

Jane and Sarah are best friends. They both entered the same painting competition. Now Jane wanted to win this competition very much indeed, but when the results were announced it was her best friend Sarah who won, not her. Jane was very sad she had not won, but she was happy for her friend who got the prize. Jane said to Sarah, "Well done, I'm so happy you won!". Jane said to her mother, "I am sad I did not win that competition!".

Q1. Is it true what Jane said to Sarah?
Q2. Is it true what Jane said to her mother?
Q3. Why does Jane say she is happy and sad at the same time?

It is Halloween, and Chris is going to a fancy-dress party. He is going as a ghost. He wears a big white sheet with eyes cut out to see through. As he walks to the party in his ghost costume, he bumps into Mr Brown. It is dark, and Mr Brown says, "Oh! Who is it?". Chris answers, "I'm a ghost Mr Brown!".

Q1. Is it true, what Chris says?
Q2. Why does Chris say this?

At school today, John was not present. He was away ill. All the rest of Ben's class were at school though. When Ben got home after school, his mother asked him, "Was everyone in your class at school today?". Ben answers, "Yes mummy".

Q1. Is it true what Ben said?
Q2. Why did Ben say that?