Automatic Facial Expression Analysis in Diagnosis and Treatment of Schizophrenia

Mina Adel Thabet Bishay

Submitted in partial fulfillment of the requirements of the Degree of
Doctor of Philosophy

Supervisor: Prof. Ioannis Patras

School of Electronic Engineering and Computer Science

Queen Mary University of London

United Kingdom

September 2019

Statement of originality

I, Mina Adel Thabet Bishay, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Mina Adel Thabet Bishay

Date: 26/09/2019

Details of collaboration and publications:

- **Bishay, Mina**, Georgios Zoumpourlis, and Ioannis Patras. "TARN: Temporal Attentive Relation Network for Few-Shot and Zero-Shot Action Recognition." Proceedings of the British Machine Vision Conference (BMVC), 2019.
- **Bishay, Mina**, Stefan Priebe, and Ioannis Patras. "Can Automatic Facial Expression Analysis Be Used for Treatment Outcome Estimation in Schizophrenia?." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.
- **Bishay, Mina**, Petar Palasek, Stefan Priebe, and Ioannis Patras. "Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis." IEEE Transactions on Affective Computing (2019).
- **Bishay, Mina**, and Ioannis Patras. "Fusing multilabel deep networks for facial action unit detection." 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE, 2017.

Abstract

Patients with schizophrenia often display impairments in the expression of emotion and speech and those are observed in their facial behaviour. Such impairments present valuable information for the psychiatrists, as they can be used for diagnosis. However, behaviour analysis is subjective in clinical settings and time-consuming in research settings. In this thesis, our aim is to develop fully-automatic methodologies for a) quantifying patient's facial behaviour, b) estimating symptom severity in schizophrenia, and c) determining whether the symptoms have improved or not by a given treatment. In the analysis, videos of professional-patient interviews of symptom assessment, that were recorded in realistic conditions, are used. This helps in moving from controlled contexts used in the literature to similar-to-real clinical settings. Firstly, an architecture is proposed for automatic facial expression analysis. The proposed architecture address the data imbalance and threshold selection problems in multilabel classification, and is trained using several datasets recorded in controlled environments. Then, the expression analysis is moved from the controlled environments to the recent in-the-wild settings, where VGG-16 networks are trained using 4 recent datasets captured in the wild. In-the-wild analysis helps in analyzing more patients and leads to better results in symptom estimation. Secondly, a deep learning approach is proposed for estimating expression-related symptoms of schizophrenia in two different assessment interviews, namely PANSS and CAINS. The proposed approach consists of Gaussian Mixture Model and Fisher Vector layers for extracting compact statistical features over the whole video interview. Experiments show promising results both on statistical analysis and symptom estimation. Finally, two methods are proposed for addressing directly the problem of treatment outcome estimation in schizophrenia – more specifically, are aimed at determining whether specific symptoms have improved or not by analysing jointly two videos of the same patient, one before and one after the treatment.

Acknowledgments

First of all, I would like to deeply thank my supervisor Prof Ioannis Patras for all the help and support I received during the past four years of my PhD, for putting me always on the right path, for being humorous and flexible, and for raising my goals and research standards.

Next, I need to thank my family for all what they have done for me, my parents, who kept calling and supporting me daily during this challenging period, uncle Onsy and auntie Suzie, for giving me incredible support in the UK and for being like parents for me, and my brother and sister-in-law, for their continuous motivation and support, and for always cheering me up with their little monkey, Fady.

Then I want to thank my lifetime friends back home (Kerollos, Joseph, Maged, Ibrahim, Andrew, Arsany, Daniel, Manas, Samuel, Tiha, Waseem, Ahmed, Tadros, Nahedj, Sherif, Beshoy, Amir, Samer, Peter), for their daily messages and calls that make me feel like I am in Egypt. Special thanks to those who came to visit me here in the UK.

The next thanks goes to my colleagues at Queen Mary University of London, the former colleagues (Petar, Juan, Aria, Marras, Wenxuan, Young, Ye, Silvia, John, Fiona, Sally, Faranak, Farzad, Zongyi, Nur), and the current ones (Tingting, Georgios, Christos, Ellie, Andrej, Camilo, Silvia, Sam, Yan, Maria, Yachi, Bilal, Yibao, Krishna) for their help, and for all the fun we have had in the previous years. I shouldn't forget to thank the people from other groups (Thomas, Shaker, Maksud, Fan) for the great time we spent together.

Contents

List of Figures

List of Tables

List of Abbreviations

AFEA	Automatic Facial Expression Analysis
AUs	Action Units
CAINS	Clinical Assessment Interview for Negative Symptoms
CNNs	Convolutional Neural Networks
DNN	Deep Neural Network
ECSI	Ethological Coding System for Interviews
FACS	Facial Action Coding System
FV	Fisher Vector
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Unit
MAE	Mean Absolute Error
PANSS	Positive and Negative Syndrome Scale
PCC	Pearson's Correlation Coefficient
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
TARN	Temporal Attentive Relation Network

CHAPTER 1

Introduction

Contents

1.1 Mental illnesses – Schizophrenia

The European Union Green Papers published in 2005, stated that mental health problems affect one in four citizens at some point during their lives and too often lead to suicide [49]. Mental illnesses are different from other illnesses, as they often affect people in their working age causing significant losses and burdens to the economic system, as well as the social, educational, and justice systems. Subsequently, improving the diagnosis and treatment of mental illnesses have become a priority within the National Health Service (NHS).

One of the severe mental illnesses is schizophrenia. Around 0.7% of the world population is affected by schizophrenia [110]. Schizophrenia affects the way a person thinks, feels, and behaves. Schizophrenia affects not only the patients, but also their families and the society as

a whole. The overall cost of schizophrenia in the UK was estimated to be £11.8 billion per year [9].

Symptoms of schizophrenia include positive and negative symptoms. Positive symptoms refer to behaviour or thoughts that are usually not seen in healthy people (e.g. hallucinations, delusions), while negative symptoms indicate a lack of normal mental functions like motivation, concentration, or/and expression (e.g. flat affect, impoverished speech). Negative symptoms are persistent [98], and have a greater effect on patients' quality of life in comparison with other symptoms [59]. Such symptoms are particularly difficult to assess and quantify [112]. Therefore, in this thesis we will focus mainly on negative symptoms of schizophrenia.

1.2 Non-verbal behaviour in schizophrenia

Patients with schizophrenia often show impairment in the expression of emotion and speech in comparison with non-patients [131] – this is manifested in their facial expression [93], vocal expression [85, 99], and expressive gestures [21, 134]. Patients can also show impairment in the non-verbal behaviour that invites social interaction during clinical and nonclinical interviews [82]. Non-verbal behaviour was found to change during interviews according to symptom severity [40, 81, 147]. For instance, patients with high symptom severity tend to avoid interaction by nodding less, smiling less, and looking less at the interviewer [40]. Such impairments present valuable information for the psychiatrists, as they can be used for assessing symptom severity. However, non-verbal behaviour is subjectively rated during clinical interviews.

Some psychiatric researches are concerned with the relation between schizophrenia and the patients' non-verbal behaviour [37, 40, 81, 134, 147]. To perform quantitative analysis, in these works the video intervals were manually annotated in terms of the patients' non-verbal behaviour and, subsequently, statistical analysis, such as calculation of the correlations of that behaviour with the severity of the symptoms was performed. However, manual annotation of

videos is a hard and time-consuming task and requires a special training. Therefore, building an architecture that detects automatically patients' behaviour could be highly beneficial in the diagnosis and research purposes.

Recently, there has been a growing interest in studying behaviour differences in groups of patients with schizophrenia and healthy controls, as well as diagnosing schizophrenia using Automatic Facial Expression Analysis (AFEA) [7, 135, 137, 142]. The reason for the interest is that AFEA allows objective and fast measurement of facial expressions and that can be valuable for both research and diagnosis. However, the datasets that are used in current works contain only a few patients (4-34 patients) and are recorded while they were performing controlled tasks (e.g. listening to life vignettes).

1.3 Facial expression analysis

Facial expressions are facial changes that manifest due to the motion of one or more of the facial muscles. Facial expressions are a type of non-verbal communication, that can identify human affect, emotions, and personality [14]. Two of the earliest works in facial expression analysis are the work done by Guillaume Duchenne in 1862 for determining which muscles in the face are responsible for the different facial expressions [42], and the work of Charles Darwin in 1872 for describing the universality of facial expression of emotion across different cultures [36]. Following these distinctive works, facial expression analysis has been an active research area in behavioural sciences.

Two main approaches are used in behavioural sciences for studying/measuring facial expressions (or non-verbal behaviour) [32]. In the first approach, behavioural scientists interpret the message communicated by a facial pattern, where the message is an emotional/cognitive state (e.g. anger, disgust, happiness). This approach is known as judgment-based approach. In the second approach, scientists describe the facial pattern in a coded way, that is, they decompose the facial pattern into subtle actions corresponding to the movements of different facial muscles. This approach is known as sign-based approach and can describe a wide range of

facial expressions.

A common sign-based approach used by psychiatrists to code patients' non-verbal behaviour in schizophrenia is the Ethological Coding System for Interviews (ECSI) [132]. ECSI includes 37 different behaviour patterns – 15 of which are facial expressions. Using ECSI in our analysis would require the availability of datasets (images or videos) that are annotated in terms of ECSI items. ECSI-annotated datasets in the literature are not publicly available, limiting subsequently the use of ECSI in our analysis. On the other hand, there is another sign-based approach that has been extensively used by behavioral scientists in many fields, named Facial Action Coding System (FACS) [44]. FACS has different combinations of facial muscle movements, that result in different facial expressions. These muscle movements are represented by Action Units (AUs). By comparing FACS and ECSI, we found that 12 out of 15 facial ECSI items are either the same or similar to AUs in FACS. Moreover, there are many FACS-annotated datasets, that are publicly available in the literature. For this reason, we turned the problem from ECSI to FACS items detection. Table 1.1 shows ECSI and FACS similarities.

Manual annotation of facial expressions (or AUs) is a very hard task as it requires hours for annotating a minute of a video. Subsequently, building an automatic and reliable architecture for AUs detection will have a great impact on many fields e.g. affect recognition, and psychological studies. In 1978, Suwa *et al.* presented an attempt for Automatic Facial Expression Analysis (AFEA) from an image sequence [125]. Following that, AFEA has received remarkable attention from the Computer Vision community – where it has been developed significantly, moving from the recognition of basic emotional expressions to the detection of subtle AUs. Moreover, the analysis moved from posed to spontaneous expressions detection. However, most of the analysis was performed on frontal or near-frontal faces, and in controlled settings. In the last few years, the focus of the researchers is directed to real-life conditions, where facial expressions are analyzed at different head poses and recording conditions (aka in the wild).

Table 1.1: The similarities between FACS and ECSI items.

No.	ECSI	ECSI describtion	FACS	FACS describtion
1	Flash	A quick raising and lowering of the eyebrows.	AU1 + AU2	Inner Brow Raiser + Outer Brow Raiser
2	Raise	The eyebrows are raised and kept up for some time.		
3	Smile	The lip corners are drawn back and up.	AU12	Lip Corner Puller
4	Lips in	The lips are drawn slightly in and pressed together.	AU28	Lip Suck
5	Mouth corners back	The corners of the mouth are drawn back but not raised as in smile.	AU14	Dimpler
6	Shut	The eyes are closed.	AU43	Eyes Closed
7	Frown	The eyebrows are drawn together and lowered at the center.	AU4	Brow Lowerer
8	Small mouth	The lip corners are brought towards each other so that the mouth looks small.	AU23	Lip Tightener
9	Wrinkle	A wrinkling of the skin on the bridge of the nose.	AU9	Nose Wrinkler
10	Yawn	The mouth opens widely, roundly and fairly slowly, closing more swiftly.	AU27	Mouth Stretch
11	Laugh	The mouth corners are drawn up and out, remaining pointed.	AU12 + AU25	Lip Corner Puller + Lips part
12	Neutral face	A face without expression and without particular muscular tension.	AU0	Neutral face

1.4 Problem definition

The main problem we wish to solve in this thesis is to develop fully-automatic AFEA-based methodologies for diagnosis and treatment of schizophrenia, that can work in real clinical settings. We divide this problem into 3 subproblems; a) quantifying patients' facial behaviour, b) diagnosing the patient, and c) determining the treatment outcome. Specifically, our first goal is to automatically detect facial behaviour/cues that can be used to assess symptom severity and the response of patients to a treatment. The second goal is to use the detected expressions/cues for the automatic diagnosis of schizophrenia, that is, estimating the severity of the expression-related symptoms. The third goal is to determine whether the symptoms have improved or not after a given treatment, by comparing patient's facial expressions in two given video interviews, one recorded before and one recorded after the treatment. In all our analysis, we use videos of symptom-assessment interviews, which were recorded in realistic conditions

either at the patients' homes or at mental health services.

1.4.1 Challenges

In the following paragraphs we will describe the main challenges of the problems we are solving; a) AFEA in schizophrenia, b) symptom severity estimation, and c) determining the treatment outcome. Following each challenge, we will describe briefly how it is going to be addressed.

a) AFEA in schizophrenia. In this thesis we focus on using FACS for analyzing patients' facial expressions (i.e. detecting patients' AUs), as FACS can represent a wide range of facial expressions. AUs detection is a challenging task in Computer Vision due to head pose variation, appearance differences, and limitations of the available datasets, i.e. lack of sufficient positive samples for certain AUs and limited number of annotated subjects. On the other hand, the symptom assessment interviews used in our analysis were recorded at variety of places, leading to a wide range of camera viewpoints and illumination levels in the recorded videos. Moreover, patients with schizophrenia tend to gaze down or away from the interviewer, or to occlude the face by different hand gestures. This makes facial expression analysis even more challenging. In order to handle with such challenges, we **first** propose an architecture for AUs detection in settings where subjects exhibit spontaneous behaviour. This architecture is trained using 4 different datasets (available in the literature by this time), so as to a) increase the size of the training set (more subjects and video frames), and b) include different recording conditions. The main challenges in implementing this architecture are the following:

- AUs are subtle and differ according to face appearance, shape, and dynamics. Subsequently, a discriminative and rich feature representation is required for detecting different AUs. We use a Deep Learning architecture that fuses information from several sources (CNNs, MLPs, B-RNNs).
- Many works in the literature train several binary classifiers for AUs detection, that is, a binary classifier is used for each AU, in order to learn AU-specific features. Sub-

sequently, the complexity and the computational cost of the whole architecture increase linearly with the number of detected AUs. In our architecture we use a single multilabel classifier for different AUs, in order to learn general AUs features, and the embedded AUs correlations.

- The number of positive examples for the different AUs vary wildly (i.e. data imbalance). This results in the biasing of the classifier towards the class with the most samples (typically the negative class). While this can be solved in binary classification problems, for example with oversampling or undersampling. In multilabel problems, balancing the data (typically the current batch) with respect to one class (AU in our case) using oversampling or undersampling will inevitably result in unbalancing it with respect to another class. The multilabel classifier used in our architecture is modified to address the data imbalance problem by dynamically adapting the cost function.

- Different AU-annotated datasets are available for the research community. Combining datasets can improve classifier performance, however it seems a hard task when a multilabel classifier is used, since not all of the datasets are annotated for the same AUs. In this work, we adapt a multilabel classifier to be trained on all datasets by back-propagating only the errors coming from the annotated AUs in each dataset. This helps in improving the training process.

The proposed architecture shows promising performance in analyzing patients' facial expressions, and achieves the state-of-the-art results in AUs detection (by this time). However, it is restricted to frontal or near-frontal views, and to a limited range of recording conditions, as the architecture is trained using datasets collected in controlled settings. To solve this limitations, we **second** move to the recent in-the-wild analysis. That is, we train Convolutional Neural Networks (CNNs) using 4 recent datasets collected in the wild, for the detection of 10 AUs. The number of positive examples in these datasets vary immensely from one AU to another (ranging approx. between 0.6k - 35k) – this results in a heavily imbalanced data problem that is hard to be solved using the previously proposed data-balancing method. So a separate

network is trained for each AU and data imbalance is solved by undersampling in each. Furthermore, as the number of training examples are limited for some AUs, we refine VGG-16 (trained on object recognition [116]) for AUs detection. We show that in-the-wild analysis helps in analyzing more patients and leads to better results in schizophrenia diagnosis.

b) Symptom severity estimation has three main challenges. First, videos of symptom-assessment interviews used in our analysis have different lengths, as the length of the interview depends on the time spent by the patient in speaking and recollection about the interview questions. On the other hand, classifiers like MLPs or SVMs work with data of fixed dimensionality. Hence, in order to regress varying-length videos, a fixed-length representation is required to be extracted from each video. Second, videos were recorded in realistic conditions, so the conventional hand-crafted features are difficult to generalize over the different videos/patients. Third, the number of patients available for training and testing is relatively limited. In order to handle such challenges we use trainable Gaussian Mixture Model and Fisher Vector layers for extracting deep and fixed statistical representation over the whole video interview – this representation is then used with a regression layer for estimating symptom severity. The proposed model has relatively limited number of trainable parameters, which helps in reducing overfitting.

c) Treatment outcome estimation is aimed at determining whether symptoms have improved or not by a given treatment – by analysing jointly two video interviews of the same patient, one before and one after the treatment. Although symptom estimation methods could be used for this purpose, by estimating the symptom level before and after treatment, and then comparing the estimated levels, they do not perform well because the change in these symptoms is typically small [112], and falls within their margin of error. So we propose two architectures for addressing directly the problem of treatment outcome estimation in schizophrenia. Both architectures exploit deep neural networks for learning differences in patients' behaviour (facial expressions) before and after treatment.

1.5 Contributions

In this section the main contributions of the thesis are listed. In the first main chapter (Chapter 3), two architectures are developed for AUs detection, the first one is trained using data captured in controlled settings while the second is trained in the wild. The main contributions of this chapter can be listed as follows:

- A Deep Learning architecture that fuses different deep models (CNNs, MLPs, B-RNNs) together is proposed. The different models learn various kinds of information/features. Specifically, CNNs and MLPs learn deep appearance and geometric features, respectively. Moreover, a Recurrent Neural Network (RNN) is added on the top of each CNN and MLP for learning temporal features in addition to the spatial ones. In all the networks, a multilabel classifier is used, this classifier at test phase simultaneously detects all AUs.

- The inherent data imbalance in multilabel problems, and in particular in AU-annotated datasets, is addressed. Specifically, the cost term associated with each AU positive example is adjusted with the ratio of negative to positive examples in the current batch and therefore control the back-propagated error. This allows us having a single architecture for detecting several AUs, as well as addressing the data imbalance problem.

- The problem of threshold selection at the output neurons at test time is addressed. In our architecture, in order to avoid threshold selection, each class is represented by two neurons, one for positive activation while the other for negative activation. During training, those output neurons are supervised with complementary information, and during testing, the maximum of the two neurons is chosen to represent the activation.

- A comparison is presented between the two proposed architectures for AUs detection, the one trained using data captured in controlled settings and the other trained in the wild. The comparison is two-fold. First, the performance of both architectures on AUs detection is compared, and then strengths and weaknesses of each architecture are high-

lighted (in Chapter 3). Second, the effect of each architecture on the performance of symptom severity estimation in schizophrenia is showed (in Chapter 4).

In the following chapter (Chapter 4), an architecture is proposed for diagnosing schizophrenia (i.e. estimating symptom severity) in settings that are similar to the ones found in clinics and hospitals. The main contributions of this chapter can be listed as follows:

- Moving from controlled environments used in the literature to similar-to-real-life settings, where professional-patient interviews of symptom assessment are analyzed. More specifically, research interviews in which symptoms were assessed in a standardised way as they should/may be assessed in real life clinical encounters, are used in the analysis. The interviews were recorded either at the patients' homes or at the premises of mental health services across the UK. Subsequently, the recorded videos have a wide range of camera viewpoints and illumination levels that are representative of the variety of settings found in clinics. In addition, interviews of 91 outpatients are used in the analysis – this is almost 3 times the highest number of patients used in other studies.

- A Deep Neural Network (DNN) architecture, called SchiNet, is proposed. SchiNet learns deep and fixed statistical representations over videos of different lengths, and then uses these representations for estimating expression-related symptoms in two different assessment interviews. More specifically, SchiNet first uses one of the developed architectures in Chapter 3 for detecting patients' facial expressions/AUs at each frame (low-level features). Then, SchiNet uses a DNN consisting of a) Gaussian Mixture Model and Fisher Vector layers for extracting a fixed statistical feature vector over the detected expressions in the whole video interview (high-level features), and b) a regression layer for estimating symptom severity. SchiNet has relatively limited number of trainable parameters – this helps in reducing overfitting when trained on the available number of patients.

- Our experimental results show three main findings. First, some automatically detected facial expressions are significantly correlated to symptoms of schizophrenia – this confirms that symptom levels of patients with schizophrenia are expressed in the degree of their impairments in expression of emotion and social interaction. Second, several symptoms in the PANSS and CAINS interviews can be estimated with a MAE less than 1 symptom level. Third, the AUs detection architecture trained in the wild leads to more significant correlations and better symptom estimation results than the one trained using data captured in controlled settings.

Finally, in the last main chapter of this thesis (Chapter 5), two Deep Learning architectures are proposed for addressing directly the problem of treatment outcome estimation in schizophrenia – more specifically, are aimed at determining whether specific symptoms have improved or not by analysing jointly two videos of the same patient, one before and one after the treatment. In both architectures, patient's facial expressions in the two videos are first detected, and then used as input to a deep neural network. To the best of our knowledge, these are the first works to address directly the problem of treatment outcome estimation in schizophrenia. The two architectures can be summarized as follows:

- Our **first** proposed architecture uses stacked RNNs for learning local and global differences in patient's behaviour (facial expressions) before and after treatment. Specifically, a Gated Recurrent Unit (GRU) is used for learning the local differences/changes in the patient's behaviour over short concatenated clips from both videos. Then, another GRU uses the clip-level features for learning global (i.e. patient-level) features, and outputs the treatment outcome, that is a binary label that encodes whether a symptom has improved or not. This architecture is called "Stacked-RNNs". Stacked-RNNs assumes that patient's videos are aligned and have equal length (videos with different lengths are clipped).

- The **second** architecture, named Temporal Attentive Relation Network (TARN), learns

to align and compare representations (i.e. videos) of variable temporal length. The architecture consists of two modules: the embedding module and the relation module. In the embedding module, a GRU is used to extract short representations/embeddings over the facial expressions detected in short clips/segments of videos. In the relation module, a segment-by-segment attention mechanism is used first to align segment embeddings from the pair of videos. Then, the aligned segments are compared. The effect of using different comparator functions is explored. Finally, the comparison outputs are aggregated using a deep neural network consisting of two fully-connected layers and an average pooling layer – this network gives as output the treatment outcome.

The two architectures have two main differences. First, TARN is trained in an end-to-end fashion, while Stacked-RNNs is trained in two steps, and consequently TARN is easier to train and test. Second, Stacked-RNNs assumes that the patient's interviews are aligned and have equal lengths – this requires clipping videos of different lengths, and losing by that possibly useful information. On the other hand, TARN uses an attention mechanism for aligning and comparing videos of different lengths, avoiding by that any information loss. Experimental results show that using attention and the entire videos in the analysis improve the performance of the treatment outcome estimation. It is worth noting that symptom estimation methods could be used for this purpose. However, they do not perform well because the change in negative symptoms is often small [112], and falls within the error margin of these methods.

1.6 Outline of the thesis

The rest of the thesis is structured as follows. In Chapter 2, we start by reviewing the related work in analysing and diagnosing mental illnesses using automatic facial expression analysis. Then, we review the state-of-the-art methods in AUs detection. In Chapter 3, we follow by introducing the developed architectures for AUs detection, first, the one trained using data captured in controlled settings, and then the other trained in the wild. Chapter 4 presents the proposed SchiNet for estimating symptom severity in schizophrenia. Moreover, in this chapter

we introduce the clinical dataset used in our analysis. In Chapter 5, we present the first two works for addressing the problem of treatment outcome estimation in schizophrenia. Finally, we draw our conclusions in Chapter 6.

CHAPTER 2

Related work

Contents

In psychiatry, a lot of research focused on studying the non-verbal behaviour of patients with mental illnesses, like schizophrenia [134, 147], depression [51, 91], and anxiety [46, 121]. In these works, the non-verbal behaviour was manually annotated by human raters. Manual annotation is a rigorous, time-consuming process. Furthermore, the non-verbal behaviour is rated subjectively during clinical assessments. For these reasons, in the last few years there has been a growing interest in the application of automatic behaviour analysis methods for analyzing patients with mental illnesses – more specifically, for a) studying patients' behaviour, b) classifying subjects (patients vs non-patients), and c) diagnosing (i.e. estimating symptom severity) of mental illnesses. In the **first** section of this chapter, we will review related work in automatic behaviour analysis for mental illnesses.

Many of the related works use Automatic Facial Expression Analysis (AFEA) for analyzing patients' non-verbal behaviour, detecting mostly either basic emotional expressions, or facial Action Units (AUs). AUs are facial muscle movements that result in different facial expressions. In this thesis we will use AUs detection methods for analyzing patients with

schizophrenia, as AUs can represent a wide range of expressions. In the **second** section of this chapter, we will describe briefly the recent works in AUs detection, and refer the reader looking for a more detailed summary to surveys/books such as [47, 94, 128].

2.1 Automatic behaviour analysis for mental illnesses

In the last years, there has been a growing interest in the application of automatic behaviour analysis methods for studying and diagnosing mental illnesses. Two of the earliest works in this area are the work done by Alvino *et al.* [7] in 2007 for studying the differences in facial behaviour between patients with schizophrenia and healthy controls, and the work done by Cohn *et al.* [33] in 2009 for classifying depression using facial and vocal expression analysis. Following these two works, automatic behaviour analysis in mental illnesses has been an active research area in Affective Computing. In the following subsections, we will describe briefly the work done in; a) schizophrenia, b) depression, c) bipolar disorder, and d) Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD). Note that different objective markers have been used in the literature for analyzing mental illnesses, like visual markers [103, 58, 27, 65], speech/audio markers [35, 53, 63], and physiological markers [22, 43, 28] – in this chapter, the visual markers will be our main focus on presenting the related work.

2.1.1 Schizophrenia

We will review the related works in schizophrenia in terms of the datasets and the AFEA methods that are used, in addition to the main objectives of these works.

Datasets. Due to the difficulty and the ethical issues in the collection and management of data depicting patients' behaviour, there are only a few datasets available in the domain of schizophrenia. Two datasets are used in a number of works; the first one is collected in a mental health centre at the University of Pennsylvania (Penn), while the second at the Hebrew University of Jerusalem (HUJI). In this thesis we refer to the former as Penn-dataset, and the

later as HUJI-dataset. The Penn-dataset consists of videos and images that were collected at two different sessions. In the first session, patients with schizophrenia and healthy controls were asked to express basic emotions at 3 different intensities. In the second session, they were recorded while listening to vignettes about a situation in their life that is presented by them before recording. Each vignette is expected to evoke 1 of 4 basic emotions; happiness, sadness, anger and fear. The number of participants in this dataset varies across different studies [7, 57, 58, 142, 143], but it is at most 28 patients and 26 controls. The HUJI-dataset is recorded while subjects (patients and healthy controls) were participating in structured interviews. During these interviews, the participants were asked emotional questions, and also shown 20 emotional images from the International Affective Picture System. This dataset has 34 patients and 33 healthy controls, and it is used in [135, 136, 137].

AFEA methods. Different methods have been used/proposed in the literature for analysing patients' facial behaviour. In [7], Alvino *et al.* detected statically emotional expressions by measuring a deformation between a neutral face and a face with expression, which was then classified using an SVM classifier. In [142], Wang *et al.* proposed the use of temporal facial information (as opposed to only static) for analysing emotional expressions. To do so, first an SVM classifier trained using geometric features was applied for estimating the probabilities of expressions at each video frame and then a sequential Bayesian estimation, with the goal of propagating probabilities throughout the video, was applied. In [57, 58], Hamm *et al.* moved from analysing basic emotions to the detection of 15 AUs at every frame of the sequence. AUs were detected by training a Gentle Adaboost classifier using geometric and texture features. A problem with those AFEA methods is that they were trained on frontal views and on evoked expressions from professional actors. Similar results are reported in other studies: For example, [24] reported that the commercial 3D facial analysis technique used for detecting 23 AUs in [135, 136, 137], has restrictions on the distance between the user and the camera as well as the working environment.

Analysis. Several studies focused on comparing a group of patients with schizophrenia to

a group of healthy controls in terms of information/features extracted from facial expression analysis investigating the existence of differences between them. In addition, correlations between these features and the flatness and inappropriateness symptoms in the SANS scale [8] were tested. Various features were extracted in these studies. In [7, 142], the average probability of 4 emotions and neutral expression were calculated. In [143], 2D geometric features and 3D curvature features were used in the comparison. In [57], features as frequency of some single and combined AUs were extracted, while in [58] information theory measures were used as features for comparing and assessing ambiguity and distinctiveness of subjects' facial expressions. Correlations were found to be significant with the flatness symptom, and insignificant with the inappropriateness symptom. Furthermore, in [136] the facial activity of patients and controls, watching a set of emotionally evocative pictures, was analysed and used for differentiating flat and incongruent affects in schizophrenia. Variance analysis over the facial activity was used to measure flatness (variance in expressions) and incongruity (relative variance in response to similar stimuli).

A few studies by Tron *et al.* [135, 137] go beyond studying the differences in behaviour between patients and healthy controls, and more specifically, use automatic analysis of facial behaviour for the classification and severity estimation of some symptoms in the PANSS scale [75] (especially the flat affect). In these studies, different features were extracted and used with a two-step SVM based algorithm for the classification and symptom severity estimation. In [135], features related to the intensity and dynamics of each AU (e.g. frequency, activation length, change ratio) were extracted, while in [137], clustering analysis was used over all AUs for extracting 3 flatness-related features; richness (number of facial clusters appeared), typicality (the similarity to a prototype), and cluster distribution (the activation frequency of different clusters). These features were calculated over short video segments, and used for training an SVM classifier for segment-level label prediction. Then, another SVM classifier was trained using the mean and standard deviation of the segment-level predictions of each video for predicting the video-level score. [135] obtained the best classification accuracy (85%) and symptom-estimation correlation (0.53) in schizophrenia.

2.1.2 Depression

The related work in depression will be summarized based on the work main objective into 3 parts; a) classification of depression (depressed vs non-depressed), b) depression severity estimation, and c) studying depression behaviour.

a) Classification of depression. The aim of of such methods is to develop automatic methodologies for classifying/detecting depression in a group of depressed and non-depressed people. In [33], Cohn *et al.* trained an SVM classifier using Active Appearance Model (AAM) [34] based features for depression detection. Specifically, frame-to-frame differences in the coefficients of each AAM shape eigenvector were used for calculating segment-based statistical features (e.g. mean and standard deviation) for classification. In that work, a dataset consisting of 107 interviews (66 depressed, 41 non-depressed) from 51 subjects were used in the analysis. In [69], patients' upper body movements and facial changes were analyzed using spatio-temporal texture features such as Space-Time Interest Points (STIP) [80] and Local Binary Pattern on Three Orthogonal Planes (LBP-TOP) [155]. Then, these features were clustered to reduce dimensionality, and used for training a classifier for depression detection. Different classifiers have been tested in the analysis like probabilistic neural network, SVM, and Restricted Boltzmann Machine (RBM). RBM trained using STIP features achieved the best performance among other classifiers. In [70], Joshi *et al.* used relative movement of 9 body parts with respect to the torso, in addition to the whole body movements (STIP features) for depression detection. In [71], the contribution of the expressions/gestures of the different upper body parts (face, head, entire upper body) in depression detection was investigated. In [70, 71], an SVM classifier was used to classify between depressed patients and healthy controls.

Psychological research showed differences in eye movements (e.g. horizontal pursuit, blink rate) between depressed and non-depressed people [1, 87, 92]. Depressed people also showed less head movements (e.g. nodding) during speech than healthy controls [48, 56]. Based on that, ALGhowinem *et al.* proposed two methods for depression detection based on the eye

and head movements [5, 6]. In [5], AAM was used to annotate 74 points on the subjects' eyes, and then these points were used for extracting statistical features for; a) looking directions, b) blinking and eye closure, and c) horizontal, vertical, and eyelid movements. In [6], AAM was used for obtaining the subjects' head pose (yaw, roll, and pitch), and then statistical features for the different looking and tilting directions were extracted. [5, 6] trained two classifiers for depression detection, a generative classifier (Gaussian mixture model) and a discriminative one (SVM). The Gaussian mixture model classifier learns to model the feature subspace that belongs to one class, while the SVM classifier learns boundaries between classes. In [2], ALGhowinem *et al.* investigated the generalisability of using the eye and head movements/features for depression detection across three datasets collected at three countries; Australia [3], Germany [139], and USA [151]. The eye activity showed good performance over the three datasets, proving its discrimination in depression detection across different cultures. In [4], statistical features representing speaking behaviour, eye activity, and head pose were fused for depression detection. This work achieved one of the best classification accuracies (88.3%) in depression detection. In [2, 4], an SVM classifier with a radial basis function was used for depression detection. These works [2, 4, 5, 6, 69, 70, 71] used a dataset consisting of 30 depressed patients and 30 healthy subjects. Video recordings in this dataset include two parts, reading sentences and an interview with the subjects.

b) Depression severity estimation. In 2013, Valstar *et al.* released a publicly-available dataset/challenge for depression severity estimation (called AVEC 2013) [139]. AVEC 2013 includes patients performing 14 different tasks, like counting from 1 to 10, reading speech, and telling a story from subjects own past. In 2014, Valstar *et al.* released AVEC 2014 [138] with a few changes from AVEC 2013. The changes include replacing a few videos, and focusing on 2 tasks out of the 14 included in AVEC 2013. AVEC 2014 has a total of 300 videos, recorded for 84 subjects by a webcam and a microphone.

Based on AVEC 2013 and AVEC 2014, many methods have been proposed for depression severity estimation. In [73], Kächele *et al.* fused audio and visual features using a hierarchical

regressor system for estimating depression. Local Phase Quantization (LPQ) was used for extracting local appearance features over the subjects' faces. These features were regressed using a hierarchy of classifiers consisting of an ensemble of sparse regressors (e-SVR) at the base, and then two stages of multilayer perceptron. In [115], two regression models were used for estimating depression. In the first model, a single regressor was trained for predicting the full scale of depression (0-63). In the second model, a binary classifier was first used for detecting the presence and absence of depression, and then a regressor was used to predict the depression score within each class. These models were trained using audio and visual (LGBP-TOP) features. In [145], Williamson *et al.* used 20 facial AUs in addition to other audio features for estimating depression. That is, the time delay correlation and covariance matrices over four separate time delays were computed over the AUs predictions. Then, Gaussian Mixture Model (GMM) and Extreme Learning Machine (ELM) predictors were trained using the different facial and audio features for predicting the depression score. Finally, the GMM and ELM predictions were fused to give the final output using weights based on the accuracy of each predictor. This paper achieved the best RMSE (8.12) and MAE (6.31), and is the winner of the AVEC 2014 challenge.

In [66], Jan *et al.* proposed a 1D Motion History Histogram (MHH) for extracting the dynamics in the facial and vocal expressions. For the facial expressions, different texture features (LBP, LPQ, EOH) were first extracted from the facial images, and then the 1D MHH was applied for extracting a single feature vector over the different texture features. The proposed 1D MHH showed better performance than MHH that was applied to raw images in [97]. In [67], Jan *et al.* explored using deep visual features in addition to the hand-crafted ones in depression analysis – more specifically, the facial images were fed to a pretrained deep network (i.e. VGG-Face or AlexNet), and then the output features at the last fully-connected layers were acquired and used along with audio features for depression severity estimation. Moreover, [67] also proposed a feature dynamic history histogram for capturing temporal movements across the extracted features. In [66, 67], two regressors, Partial Least Squares (PLS) and Linear Regression (LR) were first trained using the extracted features for

depression score prediction, and then the PLS and LR predictions were combined using a weighted sum rule to give the final score. [67] achieved better results than the AVEC 2014 challenge winner (7.43 for RMSE and 6.14 for MAE).

c) Studying depression behaviour. Some works used automatic behaviour analysis for studying the behaviour of patients with psychological disorders, like depression. In [52], Girard *et al.* studied the relationship between the non-verbal behaviour and the severity of depression. Head motion and four AUs (AU12, AU14, AU15, AU24) were chosen for the analysis, these AUs are highly correlated to smiling, sadness, contempt, and anger, respectively. In this study, patients were compared with themselves so as to keep personal attributes the same. Specifically, patients' behaviour before treatment (high severity) is compared to their behaviour after treatment (low severity). Results showed that when the severity is high, patients avoid social affiliation by showing less head motion, smiling (AU12), and sadness (AU15), and high contempt (AU14), than when the severity is low.

In [113], Scherer *et al.* used behaviour analysis for analyzing three disorders; depression, anxiety, and Post-Traumatic Stress Disorder (PTSD). Four behaviour descriptors were used in the analysis; vertical head gaze, vertical eye gaze, smile intensity, and smile duration. Scherer found that patients with these disorders are characterized by significant increase in gazing downwards. In addition, the patients have considerable decrease in smile intensity, and smile duration, compared to those with no/less symptoms. In [122], Stratou *et al.* proposed to analyse depression and PTSD behaviour in terms of gender. Behaviour descriptors used in this study include basic emotions, facial AUs, and head gesturing. Experiments show that some behaviour descriptors are impacted by depression and PTSD, and this impact has different directions for males and females. Moreover, this impact may affect one gender, while the other is not affected. For instance, depression shows increase in AU4 for males while decrease for females. Also, contempt for PTSD increases in females while there is no difference in males. This study used virtual human assessment interview so as to test participants under the same interactions and stimuli conditions.

2.1.3 Bipolar disorder

The Audio/Visual Emotion Challenge and Workshop (AVEC 2018) introduced the Bipolar Disorder (BD) challenge for classifying patients suffering from BD into one of three categories; mania, hypo-mania, and remission [108]. Based on AVEC 2018, some works used automatic behaviour analysis for BD classification [41, 126, 148, 150]. For instance, [126] surmised that sudden changes (i.e. turbulence) in feature contours manifest the erratic behaviour of patients with BD. The turbulence in features/behaviour (such as eye-gaze, head pose, AUs occurrence) trajectories was used to capture changes in movement and emotion. Then, the extracted features used for training two types of classifiers, SVM with a linear kernel, and greedy ensembles of weighted Extreme Learning Machines. In [150], Yang *et al.* used facial AUs and body-based features for BD classification. Specifically, Yang first estimated the 2D patient's body pose, and then features like displacement between patient left and right hands, moving speed of the upper body joints were extracted, and the occurrence histogram of different AUs were extracted. Moreover, [150] used a multi-stream classification scheme along with ensemble learning for classification, that is, a video session was divided into a number of segments and then for each segment a group of random forest and statistical classifiers were trained using the extracted video and audio features for BD classification. The BD dataset consists of structured interviews for 46 Turkish speaking subjects. The best unweighted average recall achieved on the test partition of the BD challenge is 57.41% [108, 126, 150].

2.1.4 ASD and ADHD

Some works used automatic behaviour analysis for analyzing Autism Spectrum Disorder (ASD) and Attention Deficit Hyperactivity Disorder (ADHD). In [65], Jaiswal *et al.* presented an architecture for the classification of ASD and ADHD in adults. Specifically, histogram-based features from the head pose, facial AUs, Kinect animation units, and questions response time, were extracted, and used for training an SVM for classification. In [23], Canavan *et al.* used features like eye gaze angle, average gaze fixation, and subject demographic information (age, gender) for training different classifiers (e.g. random forests, C4.5) for ASD classifica-

tion in children. [65, 23] achieved classification accuracies $> 90\%$. In [107], the engagement level of children with ASD in social interactions was predicted using low-level optical flow based features. In addition, [107] showed that head pose orientation is a highly discriminative descriptor in the engagement level prediction.

2.2 State-of-the-art methods in AUs detection

AUs detection has been the focus of many researchers for a long time. In this section, we will present some of the state-of-the-art methods so as to illustrate the main trends and highlight their shortcomings.

One of the critical steps in AUs detection is feature extraction. Extracted features can be divided roughly into hand-crafted [13, 78, 152, 156], and learned features [50, 55, 64, 157]. Each of these features can be further split into appearance and geometric ones. Learned features have shown better performance than hand-crafted ones across different contexts [79, 84, 11]. Recently, learned features and in particular appearance ones learned by Convolutional Neural Networks (CNNs) have been used in AUs detection. For instance, in [50] a CNN with 3 convolutional and 2 fully-connected layers was trained for detecting different AUs. In [30], Chu *et al.* used a deeper CNN consisting of 5 convolutional and 3 fully-connected layers. In [157], Zhao *et al.* replaced the conventional CNN filters by region-specific ones to capture local appearance changes at different facial regions. In spite of the good performance achieved by the appearance features, extracting deep geometric features and fusing them with deep appearance features have not been discovered yet in AUs detection. In [72], Jung *et al.* proved that better performance can be achieved in emotion recognition, when both deep appearance and geometric features are used. Also in [13], Baltruvsaitis *et al.* fused hand-crafted appearance features (i.e. HOGs) along with geometric features (i.e. non-rigid shape parameters and landmark locations) for better AUs detection performance.

The extracted features are then used for training several binary classifiers or a multilabel classifier for AUs detection. In [13, 64, 140], a binary classifier was used for each AU in order

to learn AU-specific features. Using AU-specific classifier increases linearly the complexity, and the computational cost of the whole architecture. In [50, 55], a single multilabel classifier was used for different AUs, in order to learn general AUs features, and the embedded AUs correlations. In [157], a similar multilabel classifier was used, but replacing the conventional CNN filters by region-specific ones. Although, these filters showed good performance in AUs detection, they led to a network with a large number of trainable parameters (approx. 56 million). This can easily make the network overfit when trained on limited data or subjects.

Another aspect is the domain for extracting the features, which can either be spatial or spatio-temporal domain. Most of the recent works focused on extracting features at the spatial domain [50, 55, 157]. In [64], Jaiswal and Valstar proposed to extract the short-term spatio-temporal features by using a 3D CNN, and the long-term ones by adding a bidirectional Long Short Term Memory (LSTM) to the 3D CNN. Although, the CNN-LSTM model showed good performance in extracting the spatio-temporal features, multiple single-label classifiers were trained, one for each AU, and therefore the AUs correlations were discarded. Unlike [64], [30] trained a single multi-label classifier for AUs detection using concatenated deep spatial and temporal features, that were extracted by a CNN and an LSTM.

The normalization of the features or the face images using the subject's neutral face helps in extracting more discriminative features. In [50], Ghosh *et al.* proposed to normalize the face images using the subject's mean (neutral) face. In [13], Baltrusaitis *et al.* proposed to normalize the extracted features by using features calculated from the subject's median (neutral) face. Although, subtracting the mean/median face can improve the performance significantly, the calculated mean/median face is not always the neutral face. Larger improvement can be achieved if an accurate neutral face image is fed to the network.

Different AU-annotated datasets are available for the research community. The way that these datasets are used in training and testing affects the AUs detection performance, and reflects the generalization of the classifier. In [64, 152], models were trained and tested on the same dataset. In [50], one dataset was used for training, while another for testing. In [13],

different datasets were combined for specific AUs so as to increase the number of training examples. Although combining datasets can improve the classifier performance, it seems a hard task when a multilabel classifier is used since not all of the datasets are annotated for the same AUs.

Finally, the literature shows that different methods have been proposed for AUs detection [13, 55, 50, 64, 157], differing in various aspects e.g. kind of classifier (single- or multi-label), domain for extracting features, and feature normalisation. However, most of these methods used datasets that were recorded in controlled environments, and from frontal or near-frontal views. Recently, the focus of the researchers is directed to real-life conditions, where facial expressions are analyzed at different head poses and recording conditions (aka in the wild).

2.3 Conclusion and discussion

Conclusion. In the first section of this chapter we have mentioned and briefly described some of the approaches that used automatic behaviour analysis for classifying, diagnosing, and studying mental illnesses. Generally over the different illnesses, we can first conclude that the majority of the work is focused on depression. More datasets and approaches are required for the analysis of other mental illnesses. Second, most of the conducted research used structured interviews in their analysis. Third, the classification of mental illnesses showed high performance, reaching sometimes more than 90%. However, the diagnosis (symptom severity estimation) of mental illnesses is still in an early stage from reaching the performance required in medical applications. Finally, most of the methods proposed for classifying and diagnosing mental illnesses rely on conventional hand-crafted features.

Focusing specifically on schizophrenia as it is the mental illness we are going to work on, we can first conclude that most of the conducted research focuses on studying behaviour differences between patients and healthy individuals and that only a couple of works address the classification and symptom estimation problem in schizophrenia. Second, the datasets used

in these works contain a relatively small number of patients (4-34 patients) and were recorded while the patients were performing controlled tasks, such as listening to life vignettes, or answering emotional questions. Third, the architectures used/proposed for facial expression analysis work either on frontal views or in a specific environment. Finally, all the features used in the classification and symptom estimation of schizophrenia are hand-crafted ones. By contrast, in our work we use video recordings of 91 patients in conditions that are similar to realistic symptom-assessment interviews. We also train the proposed AFEA methods either in the wild or using different datasets in order to be robust to different recording conditions. In addition, we use statistical deep features for estimating symptom severity.

In the second section of this chapter, we can conclude that many works address the problem of AUs detection as a binary classification problem, where a different model is built for each AU, and ignoring in this way informative correlations between the different AUs. Other works that pose the problem as a multilabel classification problem are faced with the inherent imbalance of the data, since the number of positive examples for the different AUs vary wildly. Moreover, combining different datasets in the training process becomes an impediment, as the datasets are annotated in terms of different AUs. In our work, we first train a single architecture with a multilabel classifier using different datasets (collected in controlled settings) for AUs detection. Then, we move to the recent in-the-wild analysis, and train VGG-16 networks for the detection of AUs at more various recording settings.

Discussion. One of the common gabs that need to be addressed across different illnesses is the use of structured interviews in the analysis. These interviews are different from the ones conducted in clinics and hospitals. Moving the analysis from controlled to real settings can raise many challenges. More specifically, interviews recorded in realistic conditions can have different recording conditions (illumination levels and camera poses), this can severely affect the behaviour analysis architectures. Moreover, interviews recorded in real scenarios can have different lengths, and classifiers like MLPs or SVMs work with features of fixed dimensionality. Hence, in order to regress varying-length videos, a fixed-length representa-

tion is required to be extracted from each video. In our work in schizophrenia we use video recordings of symptom-assessment interviews, that were recorded in similar-to-real-life settings.

Across the different illnesses, the methods that are typically used/proposed for analysing facial expressions of patients perform well primarily in a specific, controlled environment – these methods are hard to be used in real scenarios. In Chapter 3, we propose two architectures for analysing facial expressions, that are either trained in the wild or using different datasets in order to be robust to different recording conditions. First, a single architecture with a multilabel classifier is trained using 4 different datasets so as to increase the size of the training set, and include different recording conditions. In the heart of this architecture, we address the data imbalance problem in multilabel classification, and propose a novel way for selecting threshold automatically at the output neurons. Second, another architecture is trained using the recent in-the-wild datasets for the detection of AUs at more various recording conditions.

Another gab in the literature is the relatively low performance in estimating the severity of schizophrenia/depression. One of the possible reasons behind that is the hand-crafted features used in the analysis – these features are difficult to generalize over different videos/patients, and subsequently can have implications on the performance of the regression models. Hand-crafted features have shown inferior performance in comparison to learned ones and in particular those learned by Deep Neural Networks. However, training deep architectures requires a large amount of data, and the number of patients available for the analysis in this kind of problems is relatively limited, due to the difficulty and the ethical issues in the collection of data depicting patients' behaviour. Subsequently, developing deep architectures that can learn distinctive features over limited number of patients is highly needed in the field. Note that training deep architectures with a large number of parameters like CNNs over limited data can easily lead to overfitting. Also, training temporal models like RNNs over long video sequences tend to suffer from the vanishing or exploding gradients problem – hence, building and training an appropriate deep architecture is quite challenging.

In the literature, we can see that across the different illnesses researchers have focused on extracting different statistical features (e.g. average, standard deviation, etc) from the detected behaviour or behaviour-related low-level features. Statistical features have shown good performance in depression classification [4, 5, 6], depression severity estimation [66, 105, 145], and schizophrenia symptom estimation [135]. Subsequently, in Chapter 4 we propose a deep statistical-based architecture (named SchiNet) for estimating symptom severity in schizophrenia. SchiNet consists of trainable Gaussian Mixture Model and Fisher Vector layers for extracting deep and fixed statistical representation over varying-length videos. SchiNet has relatively limited number of trainable parameters.

Although many works have focused on classifying or estimating the severity of mental illnesses. To the best of our knowledge, no works have addressed the problem of treatment outcome estimation, that is, determining whether symptoms have improved or not by a given treatment. In Chapter 5, we propose two architectures for estimating the treatment outcome in schizophrenia. The first architecture uses stacked Recurrent Neural Networks (RNNs) for learning local and global differences in patient's behaviour before and after treatment. One RNN is used for learning behaviour differences over short video segments, while the other uses the segment-level features for learning global features. The second architecture consists of a similarity/relation network and an attention mechanism to align and compare videos of variable temporal length. The use of relation networks for estimating the treatment outcome is inspired by their success in data-limited problems (i.e. few-shot and zero-shot learning) [123].

CHAPTER 3

Automatic facial expression analysis

Contents

Our aim in this chapter is to automatically detect facial cues/behaviour that can be used to assess the symptom severity and the response of the patients with schizophrenia to a given treatment. In the literature, many psychiatric researches studied the patients' facial behaviour. A common framework that was used for studying patients' facial behaviour is the Ethological Coding System for Interviews (ECSI) [132]. Although, the ECSI behaviour items showed good correlations to symptom severity [40, 81, 147], there are not ECSI-annotated datasets that are publicly available in the literature – these datasets are essential for training of the automatic architectures. Other framework that has been used across many fields is the Facial

Parts of this chapter have been published in [17] and [18].

Action Coding System (FACS). Many facial expressions in ECSI and FACS are similar (see Table 1.1), in addition FACS has several annotated datasets. Hence, we focus on using FACS for analyzing facial expressions (i.e. detecting AUs). In this chapter, we detect some of the AUs that are similar to ECSI items, in addition, we explore other AUs that can be possibly meaningful in schizophrenia.

AUs detection has received significant attention from the Computer Vision community, due to the application of facial expression analysis in areas such as affect recognition and psychological studies. AU-annotated datasets in the literature can be divided roughly into two categories, one category (conventional) includes datasets consisting of videos recorded for participants in controlled environments and from the frontal or near-frontal views, while the other category (recent) includes datasets consisting of Internet images captured at different head poses and recording conditions (i.e. in-the-wild). The two categories vary in different aspects e.g. recording conditions, kind of data (static or dynamic), number of annotated subjects and frames/images, etc. In this chapter we develop two different architectures for AUs detection, one for each category. Each architecture is trained using the datasets available in the corresponding category. Moreover, we present a qualitative and quantitative comparison between the two architectures in different settings, in order to highlight strengths and weaknesses of each architecture. The two architectures will be explained, validated, and compared in detail in the following sections.

3.1 Facial expression analysis in controlled settings

In this section we propose our first architecture for AUs detection in controlled settings. The proposed architecture fuses information from several specialized Deep Neural Networks (CNNs, MLPs, B-RNNs), each of which models a different aspect of the AUs detection problem. At the core of our architecture is a method for training each of the individual deep networks as a multilabel classifier that at test phase simultaneously detects all AUs. We adopt this multilabel classifier to address the data imbalance and threshold selection problems. This

allows us to design a more general architecture that can be trained across several datasets and for the detection of many AUs. Extensive experimental results show that our approach outperforms the state of the art by a considerable margin. In the next subsections, we will describe the proposed multilabel training scheme and fusion architecture.

3.1.1 Multilabel training scheme

AUs detection can be naturally seen as a multilabel classification problem in which, at each example, one or typically more AUs are activated. Several works address the AUs detection problem as independent binary classification problems, where a different classifier is trained for each AU. However, the complexity of such an approach increases as the number of classes/AUs increases. In addition, [156] showed that using a multilabel classifier that exploits AUs relations/dependencies can outperform standard binary classifiers. More specifically, the occurrence of AUs is correlated, for instance the occurrence of brow lowerer (AU4) is correlated to tightening of the eye lids (AU7) during anger situations, and the occurrence of lip puller (AU12) is correlated to opening of the mouth (AU25) and raising of the cheeks (AU6) in happiness contexts. Moreover, the occurrence of some AUs disable the occurrence of others. For example, raising the eyebrows (AU1 and AU2) can not occur with lowering the eyebrows (AU4), and closing the eyes (AU43) can not occur with raising the eyelids (AU5). This kind of information (i.e. label dependencies) is discarded when the AUs detection problem is treated as independent binary classification problems. Subsequently, in our first architecture we use the deep multilabel classifier used in [101], in order to reduce the complexity of our architecture and learn the embedded label dependencies (or independencies). Such multilabel classifier is employed in each of our specialized models, that will be described in the next subsection.

Deep Learning architectures have many parameters that can easily overfit when trained on a limited number of subjects or data. Hence, increasing the training set size by combining different datasets can help in improving the training process, and learning more distinctive features. The main impediment for using combined datasets in training a multilabel classifier

is the unequal number of AUs annotated in these datasets. In order to solve this problem and exploit all the available datasets, each image in the used datasets is annotated in terms of 18 AUs, with a ground truth label $q \in \{0, 1, \text{NL}\}$. The AU presence is labeled by 1, AU absence by 0, and NL if the image is not annotated for this AU. The computed cost for the NL-labelled AUs is discarded, and does not take part in the average back-propagated cost. Therefore, the computed cost is only for the annotated AUs in each batch.

One of the contributions that we make in this field is addressing the problem of threshold selection in AUs detection. Typically, in order to make a binary decision on whether the AU in question is activated or not, the corresponding neuron output is thresholded either at 0.5 or, more commonly, by a threshold that is chosen based on the training set. However, different conditions (e.g. head motion, lighting effects) can affect the neuron output, and therefore using a certain threshold for all images is not the best choice. In order to overcome this problem and choose the threshold automatically, each AU i is represented by 2 neurons, one representing AU presence AU_1^i while the other representing AU absence AU_0^i. During training, the 2 neurons are supervised by complementary information, and during testing the one with the highest output is selected. Doing so allows the network to choose the threshold automatically according to the given input conditions.

Another contribution that we make in this field is a scheme that addresses the problem of data imbalance. Data imbalance is a common problem in many applications including AUs detection and results in the biasing of the classifier towards the class with the most samples. Typically, positive examples are limited - this can be tackled by duplicating the positive examples (named "Oversampling"), or removing some negative examples (named "Undersampling") [25]; however, this is only possible in a binary classification problem – in a multilabel classification problem balancing the data with respect to one AU will result in unbalancing it with respect to other AUs. In our architecture, we propose a new method for balancing the data in a multilabel classifier. For each batch in the training set, let us denote by p_i the number of the positive examples and n_i the number of negative examples for the AU i.

Then, the ratio r_i of the negative to positive examples is computed as:

$$r_i = \begin{cases} \frac{n_i}{p_i}, & \text{if } n_i \; and \; p_i \neq 0 \\ 1, & \text{otherwise,} \end{cases} \tag{3.1}$$

where the index $i \in \{1, 2, 3, ..., 2K\}$ where $2K$ is twice the number of the detected AUs, as we use 2 neurons for representing each AU. Then, we create a weight matrix M, having the same size of the output batch. In the weight matrix, we set the 0-labeled examples by 1, the NL-labeled by 0, and the 1-labeled by r_i, where r_i is given by Equation 3.1. This weight matrix is multiplied elementwise by the output cost matrix. By doing so we adjust the misclassification cost of the positive examples so as to prevent the biasing of the network towards the negative class when a few positive examples are available. We use the binary cross-entropy as a cost function. That is, the total batch cost is:

$$C = -\frac{1}{2K} \sum_{i=1}^{2K} \frac{1}{z_i} \sum_{j=1}^{bs} M_{ij}(t_{ij} \log q_{ij} + (1 - t_{ij}) \log(1 - q_{ij})) \tag{3.2}$$

$$z_i = \sum_{j=1}^{bs} M_{ij} \; , \tag{3.3}$$

where z_i denotes the sum of the weights at AU i, bs the batch size, t the target value and q is the predicted value. Figure 3.1 shows how the automatic threshold selection method formulate the ground truth of a training batch, and also shows the weight matrix $\boldsymbol{M}$ generated for balancing the data samples in this batch.

3.1.2 Fusion architecture

The proposed architecture for AUs detection consists of multiple deep networks, as depicted in Figure 3.2. Specifically, two CNNs are used for extracting appearance features, and two MLPs are used for extracting geometric features. Then, a RNN is added on the top of each of the spatial models (CNNs, MLPs) for learning the temporal dynamics of the AUs. Finally, the predictions of the different networks are fused using a linear layer – this layer picks the best weights for each AU over the different networks. In the following paragraphs we will explain the proposed architecture.

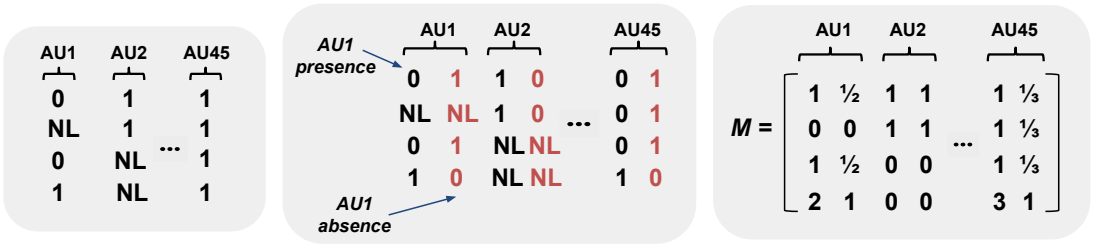


Figure 3.1: (a) Ground truth of a training batch. (b) Ground truth used when applying the automatic threshold selection method. (c) The weight matrix M generated for balancing the data.

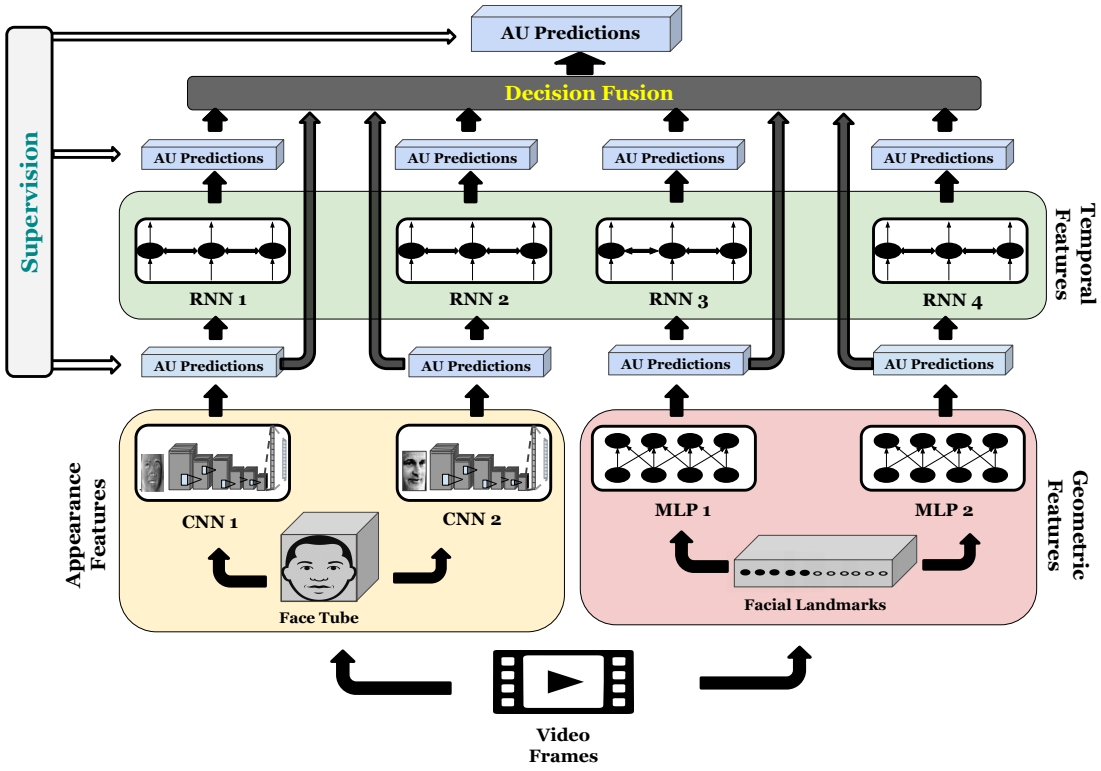


Figure 3.2: The proposed architecture.

a) Preprocessing steps. Preprocessing is crucial to ensure that a stream of aligned face images and landmarks are fed to our architecture. Preprocessing consists of 4 steps. First, we detect the subject's face using two face detectors; the OpenCV face detector, trained on frontal and profile faces, and the Zhu-Ramanan face detector [158]. We first use OpenCV due to its

(a) (b) (c) (d) (e) (f)

Figure 3.3: The preprocessing steps. (a) Input frame. (b) Detected face. (c) Detected facial landmarks. (d) Aligned face. (e) Resized face image. (f) Gray-scale face image.

fast performance, and then the Zhu-Ramanan face detector is used as a complementary model to process the failed frames. Second, we use [149] for extracting 49 facial landmarks. Third, we align the detected landmarks to a reference frame using Procrustes transform. We use in the alignment the points that are invariant to facial expressions (i.e. eye corners and nose tip). Finally, we scale the faces to a fixed resolution of 48×48, and then convert it to the gray scale. The aligned faces and landmarks are then used as inputs to the CNNs and MLPs, respectively. The complete preprocessing steps are shown in Figure 3.3.

b) Convolutional Neural Networks (CNNs). AUs detection is recently being treated as a pattern recognition problem, where one trains in a supervised manner classifiers that receive as input an image, or features extracted from it, and give at the output a set of binary labels, as many as the AUs that the method detects. In recent years the low-level feature extraction and the classifiers are replaced by CNNs [64, 157], since they have shown to learn better and more general appearance features compared to hand-crafted ones [79]. In our architecture, we use two CNNs for extracting deep appearance features. The first CNN (called CNN1) uses as input normalised face images, that is the aligned face images normalised by subtracting subject's mean face (neutral face) in the whole video. CNN1 can learn distinctive features away from the subjects' appearance differences. The other CNN (called CNN2) uses as input the aligned face images. This network works better than CNN1 when the calculated neutral face is not accurate enough. More specifically, the assumption of using the mean of all subject's frames as the neutral face is not always accurate. Figure 3.4 shows the mean face for 4 subjects,

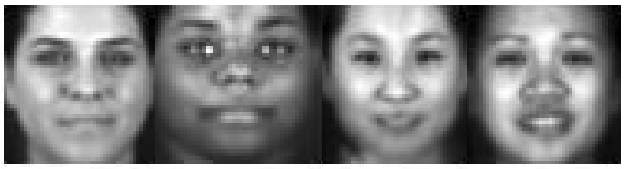


Figure 3.4: The mean faces for some subjects (selected from the BP4D dataset).

the first two images are almost neutral, while the last two are not (as the mouth is slightly opened in them). Subsequently, we use CNN2 to better classify AUs when the neutral face is not accurate enough. CNN1 and CNN2 complement each other for better AUs detection performance.

CNNs take as input face images of size 48×48. Each image is randomly cropped into 44×44 smaller sub-image, and then randomly flipped horizontally with a probability of 0.5 so as to augment the data and avoid overfitting problems. At test time, we use 44×44 sub-images cropped from the center of the aligned face images. Each CNN consists of three convolutional and one Fully Connected (FC) layers, and each convolutional layer is followed by a max-pooling layer. For the first and the second convolutional layers, we use 64 filters of size 9×9 and 5×5, respectively, and for the last convolutional layer, we use 128 filters of size 5×5. The max-pooling layers have filters of size 2×2. The FC layer consists of $2K$ sigmoid units, where K is equal to the number of the detected AUs. The activation function used in the three convolutional layers is the Rectified Linear Unit (ReLU) [100]. We use dropout for regularization [120]. We train CNN1 and CNN2 using stochastic gradient descent with 0.005 learning rate, 0.9 momentum, and 0.25 dropout. The learning rate decays (with increasing epochs) at a rate of 0.001 for CNN1 and 0.0005 for CNN2.

c) Multi-Layer Perceptron (MLP). The inspiration of using MLP along with CNN for AUs detection, comes from its success in emotion recognition in [72]. In our architecture, we use two MLPs for extracting deep geometric features. The first MLP (called MLP1) is trained using normalised facial landmarks, that is the aligned landmarks normalised by subtracting the subject's mean landmarks (neutral face landmarks) in the whole video. The other MLP

(called MLP2) is trained using facial landmarks normalized according to the method in [72]. The idea of using two MLPs is the same as with CNNs, they complement each other when the subject's neutral landmarks are not accurate enough.

The number of the detected landmarks is 49, and each landmark has two coordinates, so the length of the input feature vector is 98 (49×2). We use two hidden layers for each MLP, each consisting of 600 neurons. The output layer consists of $2K$ sigmoid units, where K is the number of the detected AUs. We use ReLU as an activation function after each hidden layer. We train MLP1 and MLP2 using stochastic gradient descent with 0.005 learning rate, 0.9 momentum, and 0.25 dropout.

d) Recurrent Neural Networks (RNNs) is a class of neural networks that is used for learning sequential information. RNNs have been extensively used in many areas such as speech recognition [54] and natural language processing [118]. In our architecture, we use the Bi-directional RNNs (B-RNNs) proposed in [114] for extracting temporal features over a sequence of frames. B-RNNs transform a sequence of inputs X to a sequence of outputs Y based on the input values, and previous and future information. Following [83], we use the ReLU function and the scaled identity initialization in our B-RNNs. At frame t, the output y_t is calculated as follows:

$$y_t = a(W^f_{out} h^f_t + W^b_{out} h^b_t + b_{out}), \quad t \in \{1, 2, ..., T\} \tag{3.4}$$

$$h^f_t = a(W^f_{in} x_t + W^f_h h^f_{t-1} + b^f_h) \tag{3.5}$$

$$h^b_t = a(W^b_{in} x_t + W^b_h h^b_{t+1} + b^b_h)\,, \tag{3.6}$$

where W^f_{out}, and W^b_{out} are the output weight matrices connecting the forward and backward hidden states to the output layer, respectively. W^f_{in}, and W^b_{in} are the input weight matrices connecting the input layer to the forward and backward hidden states, respectively. W^f_h, and W^b_h are the forward and backward hidden weight matrices, respectively. b_{out}, b^f_h, and b^b_h are

the output, forward and backward hidden bias vectors, respectively. T is the length of the video sequence. The activation function a is the sigmoid function for the output layer, and the ReLU function for the hidden layers.

In our architecture, we use four B-RNNs, one on the top of each of the spatial models (CNNs, MLPs). Specifically, B-RNNs take as input the outputs of the CNNs and MLPs over different video frames. We partition videos into segments of length 90 frames. We initialize the weights of the hidden layers by a scaled identity matrix, where 0.1 is chosen as the scale value. We train all B-RNNs using stochastic gradient descent, with a learning rate of 0.01, gradient clipping at 1.0, and batches of size 32 sequences.

e) Decision fusion. Appearance, geometric, and temporal features have varying AUs detection performances. In our architecture, we use decision fusion for combining the outputs of the eight deep networks (2 CNNs, 2 MLPs, 4 B-RNNs) – this helps in exploiting several sources of information/features, and boosting the AUs detection performance. Specifically, we combine the predictions of the different classifiers using a linear model, whose parameters are optimized with random search [16, 74]. In the random search, one weight is given for each AU/class in each network, and the final AU prediction is the weighted sum of all networks predictions. Random sampling from a uniform distribution is used to get weights between 0 and 1, and then each class weights are normalized to 1. We choose the best sampled weights based on the best F1-score. In our architecture, we initially use 25,000 iterations, and then a local random search is performed around the best weights chosen for the different classes. The weights for the local search are sampled from a Gaussian distribution with a mean equal to the best chosen weights, and standard deviation std of 0.5. The local search is repeated around the best chosen weights after every 1000 iterations, and at each time the std is decreased by a factor of 0.8, and stopped when it is smaller than 0.001.

3.1.3 Experiments and results

Datasets. In our experiments, we use four spontaneous datasets that are available in the literature (by this time); UNBC [90], DISFA [95], and FERA [140] (FERA includes two datasets: BP4D [153] and SEMAINE [96]). The UNBC dataset consists of videos recorded for patients (suffering from shoulder pain) performing some shoulder exercises. 25 subjects were involved in the study, and for each subject eight sessions were recorded. In total, 200 videos with 48,398 frames were annotated in terms of 11 AUs. The DISFA dataset contains videos recorded for subjects watching short video clips. These clips were chosen to elicit spontaneous emotions. 27 subjects were recorded in this dataset, where each subject was recorded for almost 4 minutes, giving in total approximately 130,000 frames. These frames were annotated in terms of 12 AUs. FERA 2015 challenge contains two datasets; SEMAINE and BP4D. The FERA organisers divided SEMAINE and BP4D into training, validation, and testing sets. The training and validation sets are released for researchers, while the testing set is kept sequestered by the FERA organizers, and researchers willing to participate in the challenge have to submit their codes for testing. The SEMAINE dataset was recorded to study social signals occurring during conversations. SEMAINE consists of videos recorded for 43 subjects responding to virtual humans. The 43 recordings were annotated in terms of 33 AUs, and divided into 16 for training, 15 for validation, and 12 for testing. The SEMAINE training and validation sets have in total 93,000 annotated frames. The BP4D dataset contains videos of people responding to emotion electing tasks. 61 subjects were involved in the study, where each subject was recorded during eight different tasks. FERA divided the 61 subjects into 21 training, 20 validation and 20 testing. The BP4D videos were annotated in terms of 27 AUs. The BP4D training and validation sets have in total 328 videos and approximately 146,000 annotated frames. We fused the different datasets for training our proposed architecture for the detection of 18 AUs, only AUs with sufficient number of positive examples are chosen for the analysis. The 18 AUs and their label distribution across the UNBC, DISFA, and SEMAINE and BP4D training and validation sets, are shown in Table 3.1. The reason we fused different datasets in the training is to a) increase the size of the training set (more subjects and video

Table 3.1: The label distribution of the 18 AUs used in our analysis (number of positive examples / number of negative examples) across four spontaneous datasets; UNBC [90], DISFA [95], SEMAINE [96] (training and validation sets) and BP4D [153] (training and validation sets).

AU	**UNBC**	**DISFA**	**SEMAINE**	**BP4D**
AU1	-	8778 / 122036	6503 / 86497	31043 / 115804
AU2	-	7364 / 123450	9232 / 83768	25110 / 121737
AU4	1074 / 47324	24594 / 106220	3512 / 89488	29755 / 117092
AU6	5557 / 42841	19484 / 111330	4975 / 88025	67677 / 79170
AU7	3366 / 45032	-	1801 / 91199	80617 / 66230
AU9	423 / 47975	7132 / 123682	240 / 92760	8512 / 138335
AU10	525 / 47873	-	1654 / 91346	87271 / 59576
AU12	6887 / 41511	30794 / 100020	17407 / 75593	82531 / 64316
AU14	-	-	965 / 92035	68376 / 78471
AU15	-	7862 / 122952	957 / 92043	24869 / 121978
AU17	-	12930 / 117884	2527 / 90473	50407 / 96440
AU23	-	-	1143 / 91857	24288 / 122559
AU24	-	-	3053 / 89947	22229 / 124618
AU25	2407 / 45991	46052 / 84762	16171 / 76829	-
AU26	2093 / 46305	24976 / 105838	5790 / 87210	-
AU28	-	-	1673 / 91327	5697 / 141150
AU43	2434 / 45964	-	3882 / 89118	-
AU45	-	-	15647 / 77353	-

frames), and b) include different recording conditions.

Experimental setup. In our first experiment, we use BP4D to show the effect of the proposed methods for data balancing and automatic threshold selection on a multilabel classifier. In the second experiment, we combine the FERA training set, UNBC, and DISFA for training our architecture, and use the FERA validation set for testing. In this experiment, we show in detail how each of the eight deep networks and their fusion model perform on AUs detection. In the third experiment, the code of the trained fusion model is submitted to the FERA organizers in order to be tested on the FERA testing set. FERA specifies 6 AUs on SEMAINE, and 11 AUs on BP4D for challenging. Using the FERA platform allows all participants to test their architectures in similar conditions. In the last experiment, we partition the BP4D dataset into 3 folds, and then iteratively use two folds for training and one fold for testing. The average performance over the 3 folds is compared to the results reported in [29, 86, 156, 157].

Performance metrics. We use the accuracy and F1-score for evaluating the performance

of our architecture. Accuracy is a widely used and powerful metric, but when the ratio of the negative to positive examples is large, the detection accuracy of the positive class is almost neglected. On the other hand, F1-score depends mainly on the detection performance of the positive class, but the number of the true negatives does not take a part in the computation. In what follows we report both metrics.

Results. In the first experiment, we show the effect of adding the Automatic Threshold Selection (ATS), Data Balancing (DB), and Cross-Dataset Training (CDT) to the MultiLabel Classifier (MLC). As an illustrative case we show the performance of one of the eight models (which is CNN1) on the BP4D dataset, which is divided in a 2:1 ratio of training to validation. We report the performance on the 14 AUs annotated in BP4D. When ATS is not used, the AU threshold is chosen based on the best F1-score and when CDT is selected, several datasets (i.e. UNBC, DISFA, SEMAINE, BP4D) are used in the training.

Table 3.2 summarizes the obtained results for the different settings (ATS, DB, CDT). Using only ATS seems to reduce the average F1-score compared to the simple MLC – a reason for that is that the increase in false negatives for AUs in which the positive/negative ratio is very low, where one output neuron is biased towards the more frequent class (i.e. "0") and the other output neuron is biased towards the most infrequent class (i.e. "1"). Our cost adaptation method for data balancing improves the MLC performance by 0.7% in F1-score and 0.5% in accuracy, but larger improvement is obtained when adding ATS with DB, where the F1-score is improved by approximately 3.3% (with a slight drop in accuracy). Finally, using ATS and DB with the expansion of the training set adds an additional 0.6% to the F1-score but reduces accuracy by almost 2.5% – a reason for that is that training and testing a model on data drawn from the same distribution, is more likely to perform better than a generic model that is trained and tested on data with different distribution [130]. Subsequently, we use ATS, DB, and CDT in the training of the 8 deep networks.

In order to show the importance of fusing information from several sources, and how the different networks perform on different AUs, the FERA (SEMAINE and BP4D) validation

Table 3.2: The F1-score and accuracy obtained for the different settings of the proposed multilabel classifier.

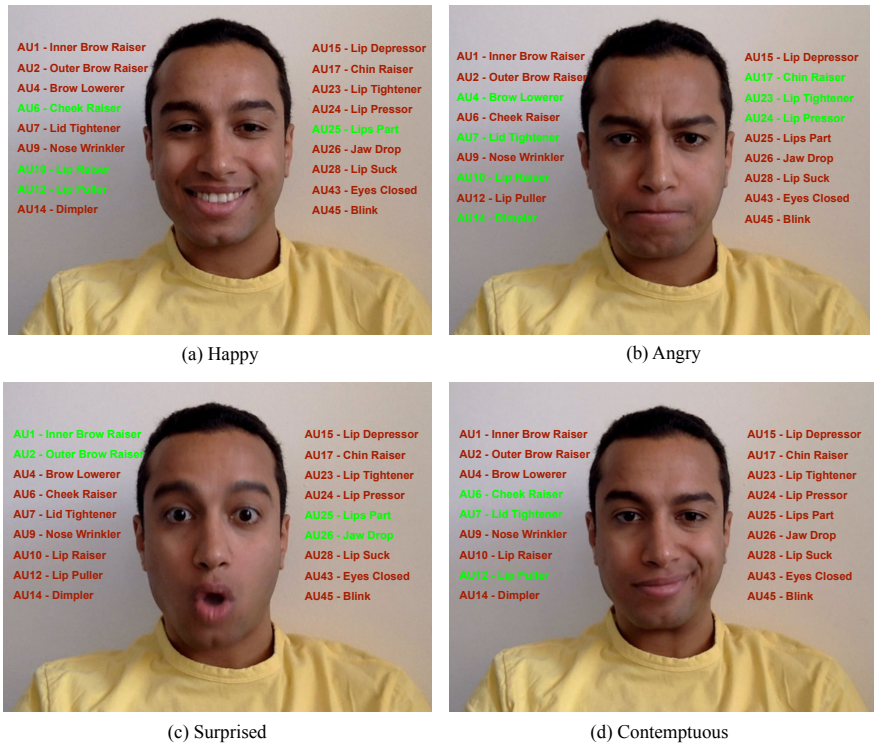
AU	MLC		ATS		DB		ATS + DB		ATS + DB + CDT	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
AU1	0.514	83.02	0.517	**84.84**	0.502	83.73	**0.551**	81.13	0.502	73.73
AU2	0.356	78.27	0.358	**82.20**	0.357	81.81	0.371	76.82	**0.403**	73.73
AU4	0.545	80.23	0.510	81.14	0.555	**81.67**	**0.556**	78.74	0.515	74.28
AU6	0.775	77.30	0.787	**79.10**	0.792	78.19	0.791	78.65	**0.805**	77.86
AU7	0.735	68.75	0.736	70.04	0.745	69.46	0.743	**70.70**	**0.763**	69.12
AU9	0.269	92.00	0.229	**93.04**	0.269	92.69	**0.351**	89.84	0.349	83.49
AU10	0.804	75.40	0.836	79.03	0.816	76.70	0.809	76.50	**0.846**	**79.50**
AU12	0.861	83.23	0.859	82.98	0.856	82.19	0.865	**83.79**	**0.868**	83.71
AU14	0.647	64.88	0.613	63.92	**0.685**	**66.31**	0.628	62.80	0.648	61.98
AU15	0.371	78.15	0.287	**79.78**	0.345	76.67	0.454	76.17	**0.465**	69.87
AU17	0.616	71.34	0.596	**74.40**	0.629	71.97	0.637	73.15	**0.656**	70.42
AU23	0.398	79.54	0.352	**82.26**	0.423	79.49	**0.465**	75.82	0.461	73.42
AU24	0.445	**83.82**	0.287	82.55	0.397	81.83	0.525	83.29	**0.562**	82.49
AU28	0.363	95.33	**0.429**	**96.70**	0.426	96.46	0.416	95.17	0.403	94.89
Avg	0.550	79.38	0.533	**80.86**	0.557	79.94	0.583	78.75	**0.589**	76.32

set is used for testing our architecture. The F1-score and accuracy obtained by the different networks, are shown in Table 3.3. By comparing the performance of the appearance (CNNs) and geometric (MLPs) features, we found that the appearance features perform better on average. However, the AUs detection performance varies over MLPs and CNNs – typically, CNNs detect better AUs that are characterized by a subtle change in the appearance (e.g. AU2, AU6, AU10, AU17), while MLPs perform better for AUs characterized by a large displacement in the landmarks' locations (e.g. AU25, AU26, AU28).

The effectiveness of the neutral face subtraction can be inferred from the good performance achieved by CNN1 in comparison to CNN2. On average, CNN1 outperforms CNN2 on both F1-score and accuracy. CNN1 works better for most of the AUs, except some of those related to the mouth area (e.g. AU15, AU17, AU24, AU25). This is due to the inaccuracy in the neutral face detection at the mouth region. Similarly, comparing the performance of MLP1 and MLP2 gives similar conclusions. The fusion of the four spatial models (CNN1, CNN2, MLP1, MLP2) leads to the second best performing model, where the F1-score is improved by approximately 3% and the accuracy by 1% compared to the best spatial model CNN1.

Table 3.3: The F1-score and accuracy obtained by the different deep networks used in the proposed architecture.

AU	**CNN1**		**CNN2**		**MLP1**		**MLP2**		**CNN1 -RNN**		**CNN2 -RNN**		**MLP1 -RNN**		**MLP2 -RNN**		**CNNs-MLPs Fusion**		**CNNs-MLPs -RNNs Fusion**	
	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
AU1	0.469	80.59	0.261	67.38	0.466	79.68	0.323	68.74	**0.489**	**82.80**	0.263	64.82	0.458	76.50	0.308	55.88	0.453	80.43	0.466	81.94
AU2	**0.448**	80.31	0.264	67.30	0.391	77.92	0.274	68.42	0.432	82.06	0.273	65.32	0.407	78.28	0.298	60.94	0.441	81.86	0.439	**82.40**
AU4	0.508	86.44	0.399	79.90	0.438	81.82	0.390	80.48	0.508	87.38	0.392	80.51	0.431	76.61	0.296	61.27	**0.558**	87.89	0.552	**88.52**
AU6	0.750	83.55	0.703	78.24	0.688	80.01	0.681	75.90	0.769	84.44	0.724	80.78	0.702	78.97	0.652	70.42	0.772	84.80	**0.775**	**85.17**
AU7	0.703	78.51	0.696	76.06	0.660	75.62	0.676	74.17	**0.718**	78.91	0.693	77.25	0.670	71.88	0.620	60.92	**0.718**	79.20	**0.718**	**79.38**
AU9	0.237	91.73	0.185	90.18	0.204	91.86	0.122	90.35	0.241	90.04	0.179	**93.68**	0.169	84.20	0.09	70.10	0.255	91.17	**0.288**	93.28
AU10	0.805	82.28	0.761	78.26	0.707	74.16	0.760	76.22	0.815	82.77	0.767	78.25	0.750	74.19	0.750	71.94	0.803	81.96	0.803	82.01
AU12	0.828	82.06	0.819	79.67	0.796	79.17	0.821	80.65	0.833	82.16	0.817	79.39	0.807	80.11	0.826	81.11	**0.841**	83.16	**0.841**	**83.17**
AU14	0.591	71.49	0.565	67.60	0.532	67.31	0.567	67.24	0.581	71.63	0.575	66.04	0.546	66.61	0.581	61.74	**0.600**	72.19	0.587	71.77
AU15	0.356	76.02	0.373	74.54	0.329	76.41	0.283	68.03	0.378	79.77	0.369	77.17	0.342	71.52	0.281	56.54	0.398	78.74	**0.409**	**80.98**
AU17	0.573	73.11	0.584	70.71	0.531	68.97	0.516	60.46	0.582	74.63	0.585	70.90	0.563	68.05	0.544	59.77	0.602	74.09	**0.603**	**74.97**
AU23	0.398	83.07	0.383	78.75	0.379	81.75	0.337	75.97	0.395	80.85	0.385	81.43	0.350	77.70	0.294	65.47	0.419	82.40	**0.427**	**84.00**
AU24	0.403	84.30	0.444	84.39	0.403	82.47	0.330	75.29	0.427	85.94	**0.458**	84.82	0.432	80.11	0.349	69.43	0.417	85.08	0.430	**85.99**
AU25	0.675	74.34	0.735	76.86	0.688	77.65	0.746	74.36	0.672	73.24	0.739	77.20	0.772	80.69	**0.780**	78.12	0.766	80.81	0.775	**81.23**
AU26	0.373	78.43	0.319	80.76	0.373	79.65	0.424	70.16	0.457	81.18	0.327	79.45	**0.477**	74.61	0.420	61.29	0.344	81.75	0.355	**81.99**
AU28	0.379	96.32	0.206	89.94	0.384	95.62	0.250	92.18	0.393	95.91	0.186	86.37	0.416	94.96	0.277	89.76	**0.509**	**97.00**	0.486	96.81
AU43	0.291	92.46	0.205	87.34	0.179	92.88	0.269	89.28	0.289	91.79	0.169	87.19	0.290	87.15	0.219	63.77	**0.353**	93.40	0.340	**93.83**
AU45	0.369	69.79	0.288	69.63	0.355	68.13	0.325	59.75	0.375	70.90	0.278	65.73	0.366	69.40	0.331	62.66	0.394	72.08	**0.398**	**72.36**
Avg	0.506	81.38	0.455	77.64	0.472	79.51	0.450	74.87	0.520	82.02	0.454	77.57	0.497	77.31	0.448	66.17	0.536	82.67	**0.539**	**83.87**



(a) Happy (b) Angry

(c) Surprised (d) Contemptuous

Figure 3.5: Results obtained by the proposed method on some videos. The 18 detected AUs are shown on the processed frames. If any of the AUs are detected, the associated text turns into green, otherwise its colour stays red.

Table 3.3 also shows the effect of adding RNN for each spatial model. The F1-score is not affected for CNN2 and MLP2, but for CNN1 and MLP1, we obtain an improvement of 1.5% and 2.5%, respectively. Adding RNN to CNN1 led to the third best model. The decision fusion of the 4 spatial models with the 4 temporal models led to the best performing model, where the F1-score is improved by 0.28% and the accuracy by 1.2% compared to the second best model (CNNs-MLPs fusion). The fusion of the 8 deep networks helps in exploiting several sources of information/features, and boosting the AUs detection performance. Figure 3.5 shows some of the detection results obtained by the proposed method on some videos depicting different emotional expressions.

The proposed fusion architecture is also tested on the FERA (BP4D and SEMAINE) testing set. Table 3.4 and table 3.5 show the obtained results on the BP4D and SEMAINE testing sets, respectively, along with other results reported in the literature [13, 55, 64, 140, 152].

Table 3.4: The F1-score obtained by the proposed method as well as other state-of-the-art methods on the BP4D testing set.

AU	B-LGBP [140]	B-Geo [140]	BCNN [55]	CDPSL [13]	DLE [152]	CNN-LSTM [64]	Proposed
AU1	0.180	0.188	**0.399**	0.260	0.261	0.280	0.349
AU2	0.159	0.185	0.346	0.250	0.167	0.280	**0.370**
AU4	0.225	0.197	0.317	0.250	0.283	0.340	**0.345**
AU6	0.671	0.645	0.718	0.730	0.729	0.700	**0.756**
AU7	0.751	0.799	0.776	**0.800**	0.785	0.780	0.776
AU10	0.799	0.801	0.797	**0.840**	0.802	0.810	0.807
AU12	0.792	0.801	0.793	0.820	0.779	0.780	**0.836**
AU14	0.666	0.720	0.681	0.720	0.625	**0.750**	0.636
AU15	0.139	0.238	0.235	0.340	**0.348**	0.200	0.344
AU17	0.245	0.311	0.368	0.330	**0.380**	0.360	0.376
AU23	0.239	0.320	0.309	0.340	**0.441**	0.410	0.426
Avg	0.442	0.473	0.522	0.516	0.508	0.520	**0.547**

Table 3.5: The F1-score obtained by the proposed method as well as other state-of-the-art methods on the SEMAINE testing set.

AU	B-LGBP [140]	B-Geo [140]	BCNN [55]	CDPSL [13]	DLE [152]	CNN-LSTM [64]	Proposed
AU2	0.755	0.569	0.372	0.410	0.655	**0.800**	0.505
AU12	0.517	0.595	0.707	0.570	**0.769**	0.740	0.702
AU17	0.066	0.091	0.067	0.200	0.215	**0.320**	0.108
AU25	0.400	0.445	0.602	0.690	0.623	**0.850**	0.810
AU28	0.009	0.250	0.040	0.260	0.251	0.330	**0.338**
AU45	0.209	0.396	0.257	0.420	0.325	**0.570**	0.451
Avg	0.326	0.391	0.341	0.425	0.481	**0.600**	0.486

We achieved the best F1-score on the BP4D dataset, and the second best on the SEMAINE dataset.

Finally, in order to compare our work with other methods in the literature, in the last experiment, a 3-fold partitioning is adopted on the combined BP4D training and validation sets, where 2 partitions are combined with UNBC, DISFA, and SEMAINE datasets for architecture training, while the remaining partition is used for testing. We report the average F1-score over the 3 runs for the 12 AUs mentioned in [157]. Our architecture is compared with other state-of-the-art methods, namely JPML [156], DRML [157], CNN-RNN [29], and EAC [86], in table 3.6. The proposed architecture outperforms the other methods on 9 out of 12 AUs, and gets the best average F1-score by a considerable margin.

Table 3.6: The F1-score obtained by the proposed method as well as other state-of-the-art methods on the 3-folded BP4D dataset.

AU	JPML [156]	DRML [157]	CNN-RNN [29]	EAC [86]	Proposed
AU1	0.326	0.364	0.314	0.390	**0.563**
AU2	0.256	0.418	0.311	0.352	**0.471**
AU4	0.374	0.430	**0.714**	0.486	0.570
AU6	0.423	0.550	0.633	0.761	**0.791**
AU7	0.505	0.670	**0.771**	0.729	0.768
AU10	0.722	0.663	0.450	0.819	**0.843**
AU12	0.741	0.658	0.826	0.862	**0.878**
AU14	0.657	0.541	**0.729**	0.588	0.662
AU15	0.381	0.332	0.340	0.375	**0.431**
AU17	0.400	0.480	0.539	0.591	**0.602**
AU23	0.304	0.317	0.386	0.359	**0.435**
AU24	0.423	0.300	0.370	0.358	**0.512**
Avg	0.459	0.483	0.532	0.559	**0.627**

3.2 Facial expression analysis in the wild

Recently, there has been a growing interest in analyzing facial expressions in real-life conditions (aka in the wild), that is at different head poses and recording conditions – this is driven with the release of new in-the-wild datasets like EmotioNet [45]. These datasets consist only of images, and vary wildly in the number of annotated samples, and in the ratio of the positive/negative examples. In order to handle with such challenges, we develop a new architecture for AUs detection in the wild, different from the one proposed in Section 3.1. The proposed architecture will be explained in the following subsections.

3.2.1 Proposed method

As the datasets released in the wild have only images, we can use here only appearance and geometric features. However, extracting meaningful landmarks-based geometric features is quite challenging, due to the wide range of head poses existing in the used facial images. Subsequently, in our architecture we extract only appearance features using a very deep CNN. The proposed architecture is trained using four datasets, for the detection of 10 AUs in the wild. The proposed architecture is shown in Figure 3.6.

Preprocessing steps. We first apply SmileNet [68] to detect the bounding box of the face

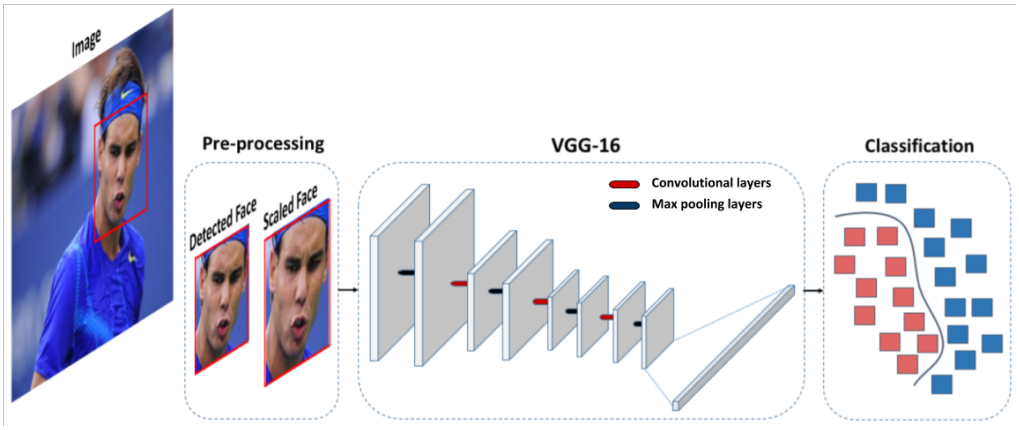


Figure 3.6: The proposed architecture for facial expression analysis in the wild.

and whether it is smiling or not. SmileNet is robust to different head poses and illumination conditions. Then, we crop and scale the detected face to a fixed resolution of 100×100. Finally, we subtract the mean RGB value of the training set from the face image – this image is used as an input to a CNN. Note that no face-registration is applied to the extracted faces.

Convolutional Neural Networks (CNNs). Following the great success achieved by AlexNet in image classification [79], CNNs have been extensively used for different Computer Vision problems in the last years. CNNs learn better appearance features than the designed ones. Furthermore, networks trained on large datasets on surrogate tasks (e.g. image classification) have been shown to perform well for feature extraction on other tasks. Motivated by this, we refine very deep CNNs (VGG-16 [116]) for the detection of 10 AUs. More specifically, we treat the problem of AUs detection, as several binary classification problems and refine separately a VGG-16 for each AU. We replace the output layer of the VGG-16 by another with a single sigmoid unit, since each network deals with a binary classification problem. We use the binary cross-entropy as the classification cost function. That is, the total batch cost is:

$$C_c(t, q) = -\frac{1}{B}\sum_{b=1}^{B}(t_b \log q_b + (1 - t_b)\log(1 - q_b)), \tag{3.7}$$

where B denotes the batch size, t the target value and q the predicted value.

Since the occurrence of AUs is correlated, some works deal with AUs detection as a multilabel classification problem [50, 55, 18]. In this architecture, we train a separate network for

each AU, because the number of positive examples vary immensely from one AU to another (ranging approx. between 0.6k - 35k) – this results in a heavily imbalanced data problem and networks that are tuned to the most populated AUs/classes. This is hard to be solved using the previously proposed data-balancing method.

In total, 11 facial expressions are analysed, ten of which are facial AUs detected using the architecture described above, and one is smile recognized using the SmileNet proposed in [68].

3.2.2 Experiments and results

Datasets. We use four datasets collected in the wild (EmotioNet [45], ExpW [154], CelebA [89], and CEW [119]), that are available in the literature (by this time), for the detection of 10 AUs – Table 3.7 shows the used datasets, as well as the detected AUs. We use different datasets in our analysis so as to detect more AUs. The facial images in these datasets were collected by searching Internet images using certain words in a variety of search engines. The collected images have different recording conditions and head poses – this improves the robustness of our model to those conditions.

The EmotioNet dataset [45] consists of 1 million images, among which 950,000 images form the training set and are automatically annotated using a developed AUs detection architecture, while 50,000 images are manually annotated and divided equally between the validation and testing sets. The EmotioNet testing set is kept sequestered by the organisers for challenging. The EmotioNet dataset consists of annotations for 12 AUs. Out of 12 annotated AUs in EmotioNet, only 7 that have sufficient number of positive examples are selected for analysis. In our analysis, only manually-annotated images in the validation set are used in the training and testing of our proposed architecture. Note that, each image in EmotioNet is annotated by "0" if the AU is inactive, "1" if active, and "999" if the AU is occluded. The Expression in-the-Wild (ExpW) dataset [154] consists of 91,793 facial images downloaded using Google search engine. Each image was manually annotated with one of the six basic emotional expressions (anger, fear, disgust, happiness, sadness and surprise) and the neutral

Table 3.7: The label distribution of the AU(s) in the EmotioNet [45] validation set, and ExpW [154], CelebA [89], and CEW [119] datasets.

Facial Expressions	**AU1 – Inner Brow Raiser**	**AU2 – Outer Brow Raiser**	**AU4 – Brow Lowerer**	**AU5 – Upper Lid Raiser**	**AU6 – Cheek Raiser**	**AU12 – Lip Corner Puller**	**AU25 – Lips Part**	**AU0 – Neutral Expression**	**AU7 – Lid Tightener**	**AU43 – Eyes Closed**
Datasets	EmotioNet							ExpW	CelebA	CEW
No. of positive examples	1268	612	2673	779	3793	6089	9643	34883	23329	1192
No. of negative examples	17655	18616	16877	18393	14620	8547	9732	56910	179270	1231

expression. CelebA is a large-scale dataset [89] consisting 10,000 identities and 200,000 facial images. Each image in CelebA is annotated with 40 face attributes. We use the annotations of only the narrow eyes attribute (which represents the lid tightener expression) for training our architecture. Finally, the Closed Eyes In The Wild (CEW) dataset consists of 2423 facial images, selected from the Labeled Face in the Wild (LFW [62]) dataset. CEW has 1192 images showing subjects with both eyes closed, and 1231 images showing subjects with eyes open. Table 3.7 shows the label distribution of the 10 AUs used in our analysis.

Training settings. We split the datasets (CEW [119], CelebA [89], EmotioNet [45], ExpW [154]) into 75% for training, 10% for validation, and 15% for testing. Many of the detected AUs have a high ratio of negative to positive examples (i.e. imbalanced data). In order to avoid the biasing of the classifier to the most frequent class (negative class), the positive and negative examples are balanced in the training set by undersampling [25]. The ExpW [154] dataset is annotated for 6 emotional expressions and the neutral expression. In order to keep the training set balanced and diverse when training for the detection of the neutral expression, negative examples equal to positive examples are drawn from all the 6 emotional expressions. For the EmotioNet dataset we trained different networks for the detection of 12 AUs, however only the networks of 7 AUs (shown in Table 3.7) show good detection performance, as these AUs have sufficient number of positive examples – those AUs are selected for further analysis.

The training set of each expression/AU is augmented with random flipping, rotation, shifting, shearing, and zooming, in order to avoid overfitting. We initialize the parameters of

Table 3.8: The classification results obtained by the proposed method in the wild on the 15% testing splits of the EmotioNet [45], ExpW [154], CelebA [89], and CEW [119] datasets.

Facial Expressions		**AU1 – Inner Brow Raiser**	**AU2 – Outer Brow Raiser**	**AU4 – Brow Lowerer**	**AU5 – Upper Lid Raiser**	**AU6 – Cheek Raiser**	**AU12 – Lip Corner Puller**	**AU25 – Lips Part**	**AU0 – Neutral Expression**	**AU7 – Lid Tightener**	**AU43 – Eyes Closed**	**Average**
Proposed method	Acc	0.941	0.869	0.903	0.857	0.880	0.908	0.919	0.731	0.855	0.980	0.884
	F1	0.459	0.319	0.632	0.304	0.716	0.897	0.912	0.718	0.526	0.977	0.646

the AUs detection networks by the parameters of the VGG-16, and then refine them using stochastic gradient descent with adaptive learning rate (RMSprop [129]), with a decay coefficient set to 0.7 and initial learning rate to 10^{-4}. Depending on the size of the training set for each AU, the batch size is set either to 64 or 128.

Results. The accuracy and F1-score obtained by the proposed method on the 15% testing splits are shown in Table 3.8. We observe that the performance is highly dependent on the number of training samples and the variance in AU-appearance. More specifically, AUs like lips part (AU25), and eyes closed (AU43) have a high value for both F1-score and accuracy, due to the relatively large number of training examples as well as the fewer differences in AU-appearance among subjects. On other AUs like brow lowerer (AU4), and lid tightener (AU7) we obtain moderately good performance due to the large variance in AU-appearance among different people. Finally, we obtain low F1-score values for the outer brow raiser (AU2) and the upper lid raiser (AU5) as the EmotioNet dataset has relatively small number of positive examples for those two classes. Figure 3.7 shows some of the detection results obtained by the proposed method on some YouTube videos.

In [15], Benitez-Quiroz *et al.* presented the EmotioNet challenge dedicated to AUs detection in the wild. The EmotioNet testing set was used as a common benchmark for comparing different methods. As the deadline for the challenge has passed and the testing set is not available for download, we couldn't compare the proposed method to other methods trained in the wild. But in general, the results we got on our testing sets are comparable to the ones reported in the challenge.

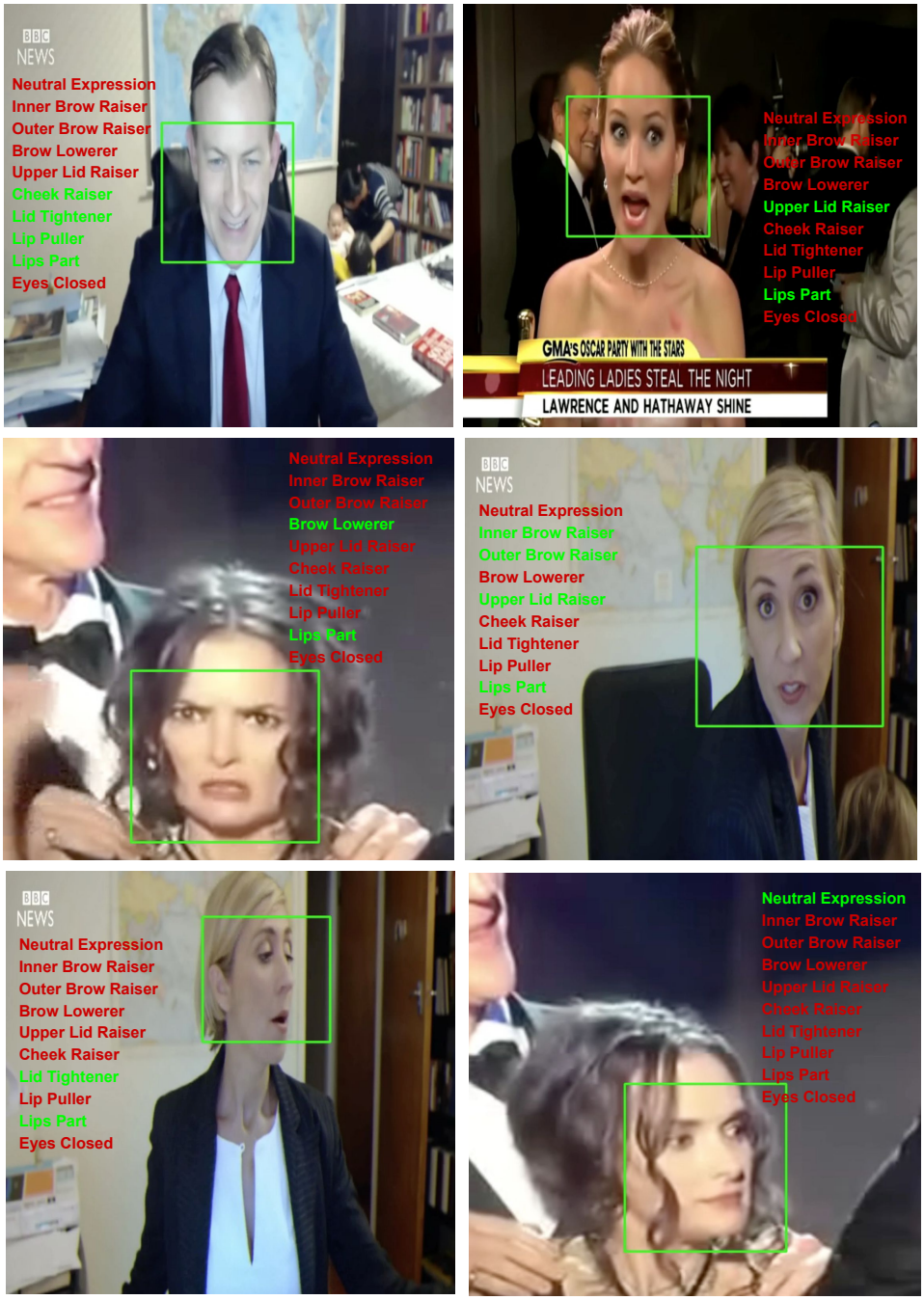


Figure 3.7: Results obtained by the proposed method in the wild on some YouTube videos. The 10 detected AUs are shown on the processed frames. If any of the AUs are detected, the associated text turns into green, otherwise its colour stays red.

3.3 Controlled versus in-the-wild facial expression analysis

In this section we compare the proposed architectures for AUs detection, the one trained in controlled settings and the other trained in the wild. The comparison is two-fold. First we test how both architectures perform on AUs detection – highlighting pros and cons of each one (in this section). Second, we show how each architecture affects the performance of symptom severity estimation in schizophrenia (in Chapter 4).

In both comparisons, only the 8 AUs that are detected by both architectures are used. Furthermore, in the comparisons we use two versions of the proposed method in Section 3.1, one version has the full architecture and is working on dynamic videos, while the other consists only of 2 out of the 8 deep networks used in the full architecture and is working on static images. More specifically, in the second version we use two spatial networks that operate on the raw facial images and the coordinates of the facial landmarks without subtraction of the mean face or landmarks (i.e. CNN2, MPL2). In what follows we will refer to the full architecture as "Full" and the simplified static version as "Static". We will first compare the performance of both architectures on AUs detection in controlled settings, and then on AUs detection in the wild.

Comparison in controlled settings. One of the main differences between the proposed architectures is the kind of extracted features. Specifically, the first method (in Section 3.1) uses various kinds of features (e.g. appearance, geometric, temporal), while the second method (in Section 3.2) uses only appearance features. This has an effect on the detection performance of each architecture. That is, moderate and intense AUs with obvious change in facial-appearance can be detected by both methods, while subtle AUs can be only detected by the first method, as they are learnt from the temporal features, or through the subtraction of the subjects' mean face/landmarks. Figure 3.8 shows a qualitative comparison between the two architectures on the FERA (BP4D and SEMAINE) validation set. Each of these images has a face with an active AU. For the first method, we use the full architecture as we are dealing here with videos. We can see that the second method can detect only AUs with high intensity, while the first

Table 3.9: The classification results obtained by the proposed architectures on the FERA validation set.

Facial Expressions		**AU1 – Inner Brow Raiser**	**AU2 – Outer Brow Raiser**	**AU4 – Brow Lowerer**	**AU6 – Cheek Raiser**	**AU12 – Lip Corner Puller**	**AU25 – Lips Part**	**AU7 – Lid Tightener**	**AU43 – Eyes Closed**	**Average**
First method (Full)	Acc	0.819	**0.824**	0.885	**0.852**	**0.794**	**0.832**	0.812	**0.938**	**0.845**
	F1	**0.466**	**0.439**	**0.552**	**0.775**	**0.718**	**0.841**	**0.775**	**0.340**	**0.613**
Second method	Acc	**0.858**	0.800	**0.895**	0.803	0.606	0.794	**0.827**	0.408	0.749
	F1	0.070	0.101	0.291	0.638	0.318	0.777	0.767	0.155	0.390

method detect both low- and high-intensity AUs. The first method fails when images having low intensity AUs are not well registered (i.e. faces are shifted or/and rotated from the reference frame). In Table 3.9, we show the performance of both architectures on the FERA validation set – we observe that on average over the 8 AUs, the first method outperforms the second method by a large margin. This considerable difference in performance is mainly due to two reasons. First, the second method is trained using facial images with high-intensity AUs (i.e. apex), while FERA has a lot of images with subtle AUs. Second, the second method uses only appearance features for AUs detection.

Comparison in the wild. The second method can work at different head poses and recording conditions, as it is trained using images captured in the wild – in contrast to the first method which is trained using datasets recorded from frontal or near-frontal views, and in specific recording conditions. Figure 3.9 shows another qualitative comparison between the two architectures on different facial images drawn from the testing splits of the in-the-wild datasets (i.e. EmotioNet, CEW, and CelebA). Each of these images has a face with an active AU. Note that here we use the Static version of the first method, as we are dealing with images. We can see that Static version of the first method performs well mainly for frontal or near-frontal faces, while the second method can detect AUs at several head poses and illumination levels. However, the second method fails when the AUs are subtle, or when the faces are captured under too dark or bright illumination conditions. Furthermore, in Table 3.10 we show the performance of both architectures on the testing splits – we observe that on average over the

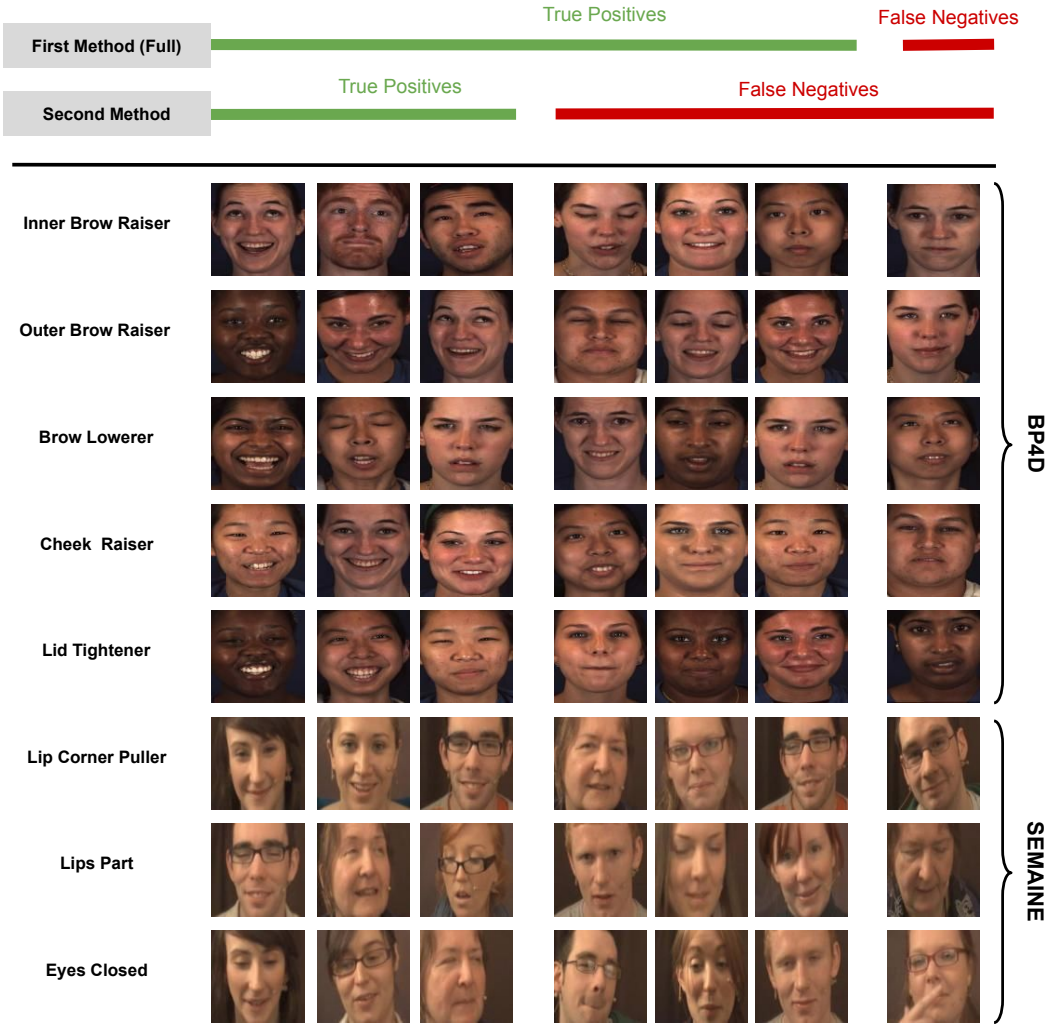


Figure 3.8: Qualitative comparison between the proposed architectures on the FERA validation set. Each row shows the positive examples of a certain AU. The true positives and false negatives achieved by each method are shown on the top part of the figure. The Full version of the first method shows better performance in detecting different levels of intensity of AUs, compared to the second method.

8 AUs, the second method outperforms the first method by a large margin. This considerable difference in performance is mainly due to two reasons. First, the first method is trained using facial images captured in controlled environments, and with a limited variation in the head pose. Second, only 2 out of the 8 deep networks in the first method are used for AUs detection. The first method showed that the full architecture can achieve better performance than both single and combined networks.

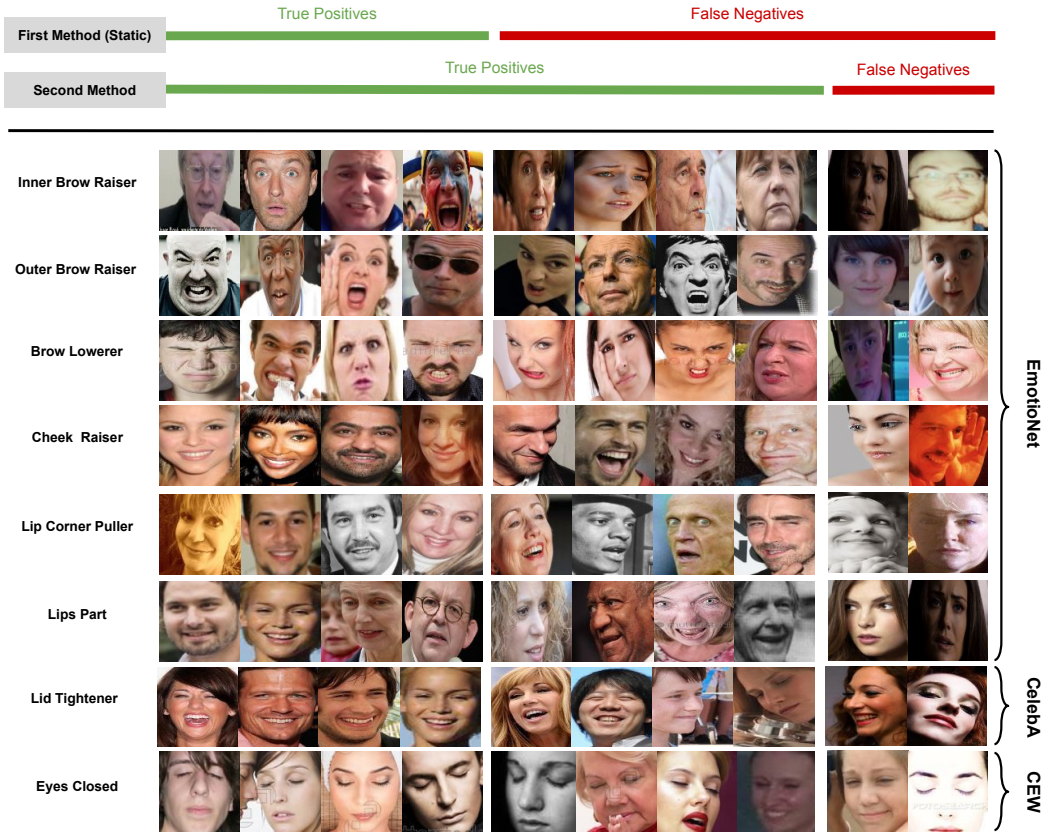


Figure 3.9: Qualitative comparison between the proposed architectures on the EmotioNet [45], CelebA [89], and CEW [119] testing splits. Each row shows the positive examples of a certain AU. The true positives and false negatives achieved by each method are shown on the top part of the figure. The second method shows better performance in detecting AUs at several head poses and illumination conditions, compared to the Static version of the first method.

Table 3.10: The classification results obtained by the proposed architectures on the EmotioNet [45], CelebA [89], and CEW [119] testing splits.

Facial Expressions		**AU1 – Inner Brow Raiser**	**AU2 – Outer Brow Raiser**	**AU4 – Brow Lowerer**	**AU6 – Cheek Raiser**	**AU12 – Lip Corner Puller**	**AU25 – Lips Part**	**AU7 – Lid Tightener**	**AU43 – Eyes Closed**	**Average**
First method (Static)	Acc	0.679	0.669	0.752	0.806	0.823	0.708	0.670	0.617	0.716
	F1	0.166	0.130	0.354	0.544	0.792	0.697	0.268	0.422	0.422
Second method	Acc	**0.941**	**0.869**	**0.903**	**0.880**	**0.908**	**0.919**	**0.855**	**0.980**	**0.907**
	F1	**0.459**	**0.319**	**0.632**	**0.716**	**0.897**	**0.912**	**0.526**	**0.977**	**0.680**

The preceding qualitative and quantitative comparisons show the strengths and weaknesses of each of the proposed methods for AUs detection. Specifically, the first method can detect

AUs at different levels of intensity, but it is working mainly on frontal or near-frontal faces and in controlled settings, while the second method detects only moderate/intense AUs (with obvious facial change), but at different head poses and recording conditions. We believe that better AUs detection can be achieved if temporal datasets, annotated for AUs in the wild, became available to the research community. In the next chapter we will test how each architecture affects the performance of symptom severity estimation in schizophrenia.

3.4 Conclusion

In this chapter we have developed two different architectures for facial expression analysis (AUs detection), one working in controlled settings while the other is working in the wild. In the first architecture, we fused different deep models (CNNs, MLPs, B-RNNs) together, in order to capture deep appearance, geometric, and temporal features. In the core of our architecture, we proposed a novel method for addressing the data imbalance problem in multilabel classification without adding any extra computational cost, in addition to a novel way for selecting threshold automatically at the output neurons. Moreover, we trained our architecture across several datasets (with unequal number of annotated AUs). This allowed us to design a more general architecture for the detection of many AUs. Experimental results show that the first method achieved the state-of-the-art results on the BP4D dataset, and outperformed other works in the literature by a significant margin.

As the datasets in the wild have only images, and vary wildly in the number of annotated samples, and in the ratio of the positive/negative examples – we developed another architecture for AUs detection in the wild. In this architecture, we refined a deep pretrained CNN (VGG-16) for AUs detection. Experiments show promising detection results at different head poses and recording conditions. In the end of this chapter, we presented a qualitative and quantitative comparison between the proposed architectures in different settings (controlled and in-the-wild), showing the pros and cons of each architecture.

CHAPTER 4

Symptom severity estimation in schizophrenia

Contents

Patients with schizophrenia often display impairments in the expression of emotion and speech and those are observed in their facial behaviour. Automatic analysis of patients' facial expressions that is aimed at estimating symptoms of schizophrenia has received attention recently. However, the datasets that are typically used for training and evaluating the developed methods, contain only a small number of patients (4-34) and are recorded while the subjects were performing controlled tasks such as listening to life vignettes, or answering emotional questions. Furthermore, the methods proposed up to now for estimating symptom severity in schizophrenia rely on conventional hand-crafted features [135, 137]. Across different con-

Parts of this chapter have been published in [17].

texts, hand-crafted features have shown inferior performance in comparison to learned ones and in particular those learned by Deep Neural Networks [79, 84, 11].

Our first contribution in this chapter is that we move from controlled environments to similar-to-real-life settings and use videos of professional-patient interviews, in which symptoms were assessed in a standardised way as they should/may be assessed in real life clinical encounters. The interviews involve a selection of patients with negative symptoms – such symptoms are particularly difficult to assess and quantify [112]. The interviews were recorded either at the patients' homes or at the premises of mental health services across the UK. Subsequently, the collected videos have a wide range of camera viewpoints and illumination levels that are representative of the variety of settings found in clinics. In addition, we automatically analyse the facial behaviour of 91 outpatients – this is almost 3 times the highest number of patients used in other studies.

The second contribution is that we propose a novel Deep Neural Network (DNN) architecture, called SchiNet, that first uses one of the developed architectures in the previous chapter for detecting patients' facial expressions/AUs at each frame (low-level features). Then, uses a DNN consisting of a) Gaussian Mixture Model (GMM) and Fisher Vector (FV) layers for extracting a compact statistical feature vector over the detected expressions in the whole video interview (high-level features), and b) a regression layer for estimating symptom severity. The GMM, the FV and the regression layer are trained in an end-to-end fashion. The proposed architecture has relatively limited number of trainable parameters, which helps in reducing overfitting.

The proposed SchiNet has been trained in a patient-independent manner to predict expression-related symptoms from two commonly-used assessment interviews; Positive and Negative Syndrome Scale (PANSS) [75], and Clinical Assessment Interview for Negative Symptoms (CAINS) [61]. Experimental results show that analyzing facial expressions in the wild delivers better performance on symptom severity estimation in comparison to the analysis in controlled settings. Furthermore, we show that high and statistically significant correlations

between the detected expressions and the severity of several symptoms in both the PANSS and CAINS can be obtained, and that the proposed network for estimating symptom severity delivers promising results.

This chapter is organized as follows: In Section 4.1, we introduce the clinical dataset of schizophrenia that we used in our analysis. In Section 4.2, we present the proposed SchiNet for estimating symptoms of schizophrenia. Finally, we report the experimental results and conclude the paper in Section 4.3 and Section 4.4, respectively.

4.1 Clinical dataset of schizophrenia

In this chapter we use a dataset called "NESS", that was collected for studying the effectiveness of group body psychotherapy on negative symptoms of schizophrenia [106]. The reason we use the NESS trial is that it was recorded in realistic conditions and in settings that are similar to the ones found in clinics and hospitals. The participants in the NESS trial were recruited from mental health services at four different places in the UK; East London, South London, Liverpool, and Manchester. In total, 275 participants were included in this study. Participants aged between 18-65, and they had a total negative symptoms score $\geq$ 18 on the PANSS interview, that is, the study focused on patients with negative symptoms. Those symptoms are typically difficult to assess and quantify [112].

The participants were assessed at three different stages throughout the study; BaseLine (BL) – before the start of the treatment, End of Treatment (EoT) – after completing 20 session of group body psychotherapy, and 6 Months Follow-Up $(6MFU)$ – 6 months after the end of treatment. Each assessment interview lasted between 40 and 120 minutes, depending on the time spent by patients in speaking and recollection about the interview questions. The patients were assessed at the interview in terms of PANSS [75] including negative, positive and general psychopathology symptoms, and CAINS [61] including experience-related and expression symptoms. In addition, other scales related to depression, quality of life and client satisfaction for patients with schizophrenia were also assessed. The interviews were completed

Table 4.1: The distribution of the labels for the CAINS expression symptoms.

Scale / **CAINS Symptoms**	**0**	**1**	**2**	**3**	**4**
EXP - Facial Expression	4	16	41	41	8
EXP - Vocal Expression	6	27	42	26	9
EXP - Expressive Gestures	10	21	27	40	12
EXP - Quantity of Speech	30	32	27	17	4

Table 4.2: The distribution of the labels for the expression-related PANSS negative symptoms.

Scale / **PANSS Symptoms**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
NEG - Flat Affect	4	8	25	39	24	8	2
NEG - Poor Rapport	10	10	40	31	13	6	-
NEG - Lack of Spontaneity and Flow of Conversation	22	22	29	11	21	5	-

in a standardised way by researchers/psychologists as they should/may be done in real life clinical encounters.

Only the assessment of the PANSS and CAINS were video-recorded from the whole interview. Most of the videos were recorded at 25 frames/s and at a resolution of 1920×1080. Out of the 275 patients, 110 accepted to be recorded at BL, 93 at EoT, and 69 at $6MFU$. Since the focus of this chapter is building a model that estimates the symptom severity for unseen patients (i.e. a generic model), only the 110 patients recorded at the BL session are used in our analysis. The average length of the recorded BL interviews is 41 minutes. The distribution of the labels for the expression-related symptoms in the PANSS and CAINS scales (across the 110 BL patients), is shown in Table 4.1 and 4.2, respectively. Note that each symptom in the PANSS scale is rated between 1 (absent) and 7 (extreme), and each CAINS symptoms has a value between 0 and 4 (0=no impairment and 4=severe impairment). More information about the dataset can be found in [106].

4.2 Proposed architecture

4.2.1 Overview

In this section we present a deep architecture, named SchiNet, for estimating symptom severity in schizophrenia from videos depicting the non-verbal behaviour of patients. Figure 4.1(a) shows an overview of the system. SchiNet takes as input a video interview for patient symptom assessment and gives as output the estimated values of expression-related symptoms and the total scale/symptoms score. Intermediate results include detection of facial expressions at frame level and statistical representations of their activations in the whole image sequence.

SchiNet performs the analysis in 4 stages; preprocessing, low-level feature extraction at frame level, high-level feature extraction at video level and symptoms regression. At the first stage, we detect the patients' faces in the video frames using a body detector [88] and a robust face detector. At the second stage, the face regions are cropped and passed to one of the developed architectures in the previous chapter for AUs detection. Encoding the patients' facial behaviour at each frame is considered as the first/low-level feature extraction. At the third stage, a Gaussian Mixture Model (GMM) and a Fisher Vector (FV) layer are used to represent the patient facial behaviour over the whole video by a compact feature vector (i.e. FV representation). The FV representation is considered as the second/high-level feature extraction. Finally, the FV is fed to two fully-connected layers for estimating the symptoms and the total score.

The training of the SchiNet is done in 3 stages, as shown in Figure 4.1(b). At each stage, a different cost is optimised. In the first stage, the network that extracts video-based representations is trained in an unsupervised manner, taking as input the sequence of the outputs of the AUs detection architecture when applied to the professional-patient interviews. More specifically, the distribution of the AUs probabilities in a video is modelled using a GMM that is implemented as a network layer. Then, the estimated GMM parameters are used to extract a FV representation for the whole video. In the second stage the FV representations are used

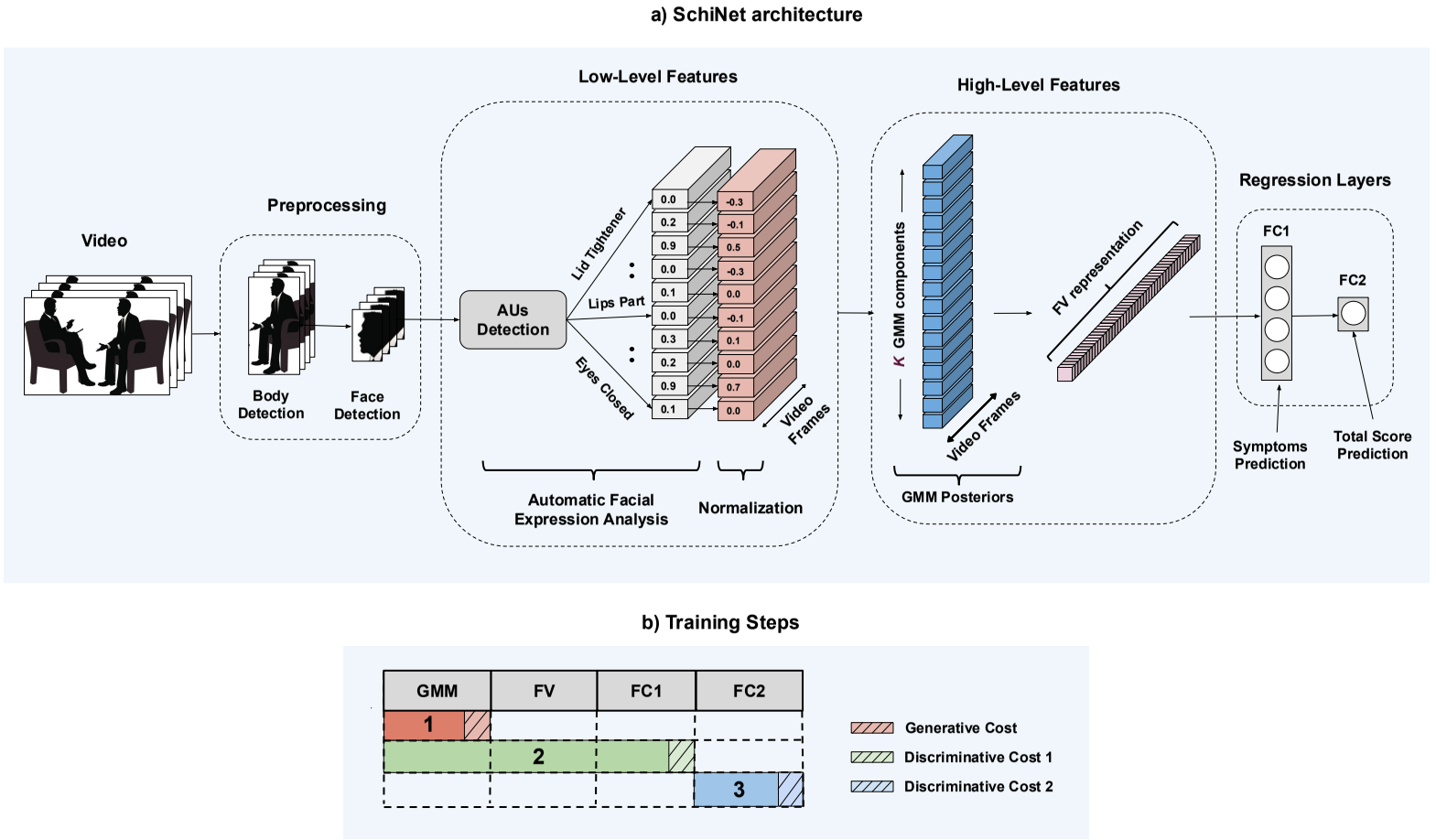


Figure 4.1: (a) The proposed SchiNet for symptom severity estimation in schizophrenia. The input is a recorded video interview of a patient during his/her symptom assessment, and the outputs are the estimated values for the expression-related symptoms and the total scale/symptoms score. Feature extraction is done over two stages, first, the video is encoded by patient facial expressions, and then a compact statistical feature vector is extracted over the encoded expressions. (b) The training stages of the SchiNet.

as input to a regression layer that estimates symptoms of schizophrenia (flat affect, poor rapport, and lack of spontaneity and flow of conversation symptoms in the case of PANSS and 4 Expression symptoms in the case of CAINS). Following [102], we refine the GMM, the FV and the first regression layer in an end-to-end fashion using a discriminative cost. Finally, in the third stage, we train a second regression layer that takes as input the individual symptom scores and estimates the total scale/symptoms score. The SchiNet architecture as well as the training stages are explained in detail in the following subsections.

4.2.2 Preprocessing steps

In order to process each video in the NESS dataset, we first extract the region of interest (i.e. the patient's face) at each frame. First, we detect the patient's body at each frame using the Single Shot Detector (SSD) proposed in [88]. We then extend the detected body-bounding box by a factor of 1.2 to ensure that the whole head is included, and then, within the resulting region, we apply the preprocessing steps (including face detection, scaling, and face registration for the method in Section 3.1) of the AUs detection architecture chosen for the analysis.

Despite the robustness of the face detection, it still fails in some videos due to the position of the camera. In those cases, not only the face is sometimes not detected but, even in the cases that it is, it is hard to be further analysed in terms of the facial expressions. For this reason, we consider only the videos in which we can successfully detect the faces (or the faces and landmarks) in more than 90% of the frames. As the preprocessing steps vary according to the used AUs detection architecture, out of the total 110 that participated in the baseline session, we retain the videos of 74 patients when the first method (in Section 3.1) is used, and 91 patients when the second method (in Section 3.2) is used.

4.2.3 Low-level feature extraction

In the second stage of the proposed method, one of the developed architectures in the previous chapter is used to code the patients' facial behaviour in terms of the Facial Action Coding System (FACS) [44], that is, detect the activation of facial AUs, the absence of which is

expected to be informative in the assessment of negative symptoms of schizophrenia. FACS has been extensively used for facial expression analysis in different contexts [141, 146, 31, 94]. Using the first AUs detection method (in Section 3.1) results in an 18-dimensional feature vector for each frame, while the second method (in Section 3.2) in an 11-dimensional vector, where each dimension represents the probability of one of the detected AUs.

Some patients in the NESS dataset have part of their faces occluded by wearable items e.g. have their eyes occluded by sunglasses or thick eyeglasses, or their eyebrows covered by a beanie hat. This results in wrong detection of the AU related to the occluded area – typically we observe false positive activations. In order to prevent these false detections from affecting the subsequent analysis steps, for each patient/video, the mean activation over each AU is calculated and subtracted from the activations of the AU in question.

4.2.4 High-level feature extraction

In section 4.2.3, we extracted frame level representations, i.e. at each frame t of the sequence we extracted a vector $\boldsymbol{x_t} \in R^M$, containing the probability of the occurrence of M facial AUs. In this section, we represent the set of vectors that are extracted for the whole video using a Fisher Vector (FV) representation. The FV representation is extracted by two custom DNN layers – the first layer learns a Gaussian Mixture Model and the second layer extracts the FV representation. The first layer is first trained using a generative cost, and then both layers are refined using a discriminative cost.

We first train a **Gaussian Mixture Model (GMM)** to model the distribution of the normalized AUs probabilities $\boldsymbol{x} \in R^M$ using a weighted sum of K Gaussian distributions [111]. Clearly, the distribution is over the set of $\boldsymbol{x}$ that are extracted over the whole training dataset, one $\boldsymbol{x}$ for every frame of each sequence. In this context, each GMM component would represent a commonly occurring combination of facial AUs. The GMM is expressed as:

$$u_\lambda(\boldsymbol{x}) = \sum_{k=1}^{K} w_k u_k(\boldsymbol{x}), \tag{4.1}$$

where w_k is the weight component of the k-th Gaussian distribution $u_k(\boldsymbol{x})$. $u_k(\boldsymbol{x})$ is defined as:

$$u_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{M}{2}} |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right). \tag{4.2}$$

Each Gaussian $u_k(\boldsymbol{x})$ has three parameters associated to it, namely the weight component w_k, the mean vector $\boldsymbol{\mu}_k$, and the covariance matrix $\boldsymbol{\Sigma}_k$. The responsibility of each Gaussian component $u_k(\boldsymbol{x})$ in generating the input feature sample $\boldsymbol{x}_t$, is called k-th posterior, and is given by:

$$\gamma_t(k) = \frac{w_k u_k(\boldsymbol{x}_t)}{\sum_l^K w_l u_l(\boldsymbol{x}_t)}. \tag{4.3}$$

In this work we follow [102], and implement the GMM as a neural network layer, that during training given a set of $\boldsymbol{x}$ learns the parameters of the GMM and during testing given an $\boldsymbol{x}$ produces K GMM posteriors $\{\gamma_t(k), k = 1, ..., K\}$ at its output (see Figure 4.1(a)). The GMM layer is first trained in unsupervised way using the Expectation-Maximization (EM) algorithm [38], that is, by minimizing the negative log likelihood (i.e. the generative cost) of the complete training data.

Once the parameters of the GMM are learned, we then represent a professional-patient video interview using a **Fisher Vector (FV) representation** – more specifically, we represent the set of low-level features, i.e. the set of vectors $\boldsymbol{x_t}$ extracted at each frame of the video in question, by a single high-dimensional vector (the Fisher Vector). The later describes how the GMM parameters should change in order to better represent the distribution of the new set of features [111], and is formed by stacking in a vector the gradients of the posteriors with respect to the GMM parameters; w_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$. Formally:

$$\mathscr{G}_\lambda^X = \left(\mathscr{G}_{w_1}^X, \ldots, \mathscr{G}_{w_K}^X, \mathscr{G}_{\mu_1}^{X'} \ldots, \mathscr{G}_{\mu_K}^{X'}, \mathscr{G}_{\sigma_1}^{X'} \ldots, \mathscr{G}_{\sigma_K}^{X'}\right)', \tag{4.4}$$

where the gradient vectors $\mathscr{G}_{w_k}^X$, $\mathscr{G}_{\mu_k}^X$, and $\mathscr{G}_{\sigma_k}^X$ are calculated as follows:

$$\mathscr{G}_{w_k}^X = \left(S_k^0 - T w_k\right) / \sqrt{w_k}, \tag{4.5}$$

$$\mathscr{G}_{\mu_k}^X = \left(\boldsymbol{S}_k^1 - \boldsymbol{\mu}_k S_k^0\right) / \left(\sqrt{w_k}\boldsymbol{\sigma}_k\right), \tag{4.6}$$

$$\mathscr{G}_{\sigma_k}^X = \left(\boldsymbol{S}_k^2 - 2\boldsymbol{\mu}_k \boldsymbol{S}_k^1 + (\boldsymbol{\mu}_k^2 - \boldsymbol{\sigma}_k^2) S_k^0\right) / \left(\sqrt{2w_k}\boldsymbol{\sigma}_k^2\right), \tag{4.7}$$

where S_k^0, $\boldsymbol{S}_k^1$, and $\boldsymbol{S}_k^2$ denote the 0-order, 1st-order, and 2nd-order GMM statistics, respectively, and are defined as:

$$\boldsymbol{S}_k^0 = \sum_{t=1}^{T} \gamma_t(k), \tag{4.8}$$

$$\boldsymbol{S}_k^1 = \sum_{t=1}^{T} \gamma_t(k)\boldsymbol{x}_t, \tag{4.9}$$

and

$$\boldsymbol{S}_k^2 = \sum_{t=1}^{T} \gamma_t(k)\boldsymbol{x}_t^2, \tag{4.10}$$

where $\gamma_t(k)$ is the k-th posterior, and T is the number of local descriptors which in our case is the video length. Following [111], the extracted FV is normalized using both power normalization, and L2 normalization.

In [102], the FV descriptor is implemented as a neural network layer, taking as input both the GMM posteriors and VGG features, and giving as output the FV. The FV layer is used also in this work, but replacing the VGG features by the normalized probabilities of the detected AUs. The layer output or the FV has a length of $K(2M + 1)$, where K is the number of GMM components and M is the feature dimensionality, which in our case is the number of the detected AUs. Note that the length of the FV does not depend on the length of the video.

Comparing the dimensionality of the low-level features (circa $500k$ for a 30-min video with 25 f/s and 11 detected AUs) to the FV dimensionality (368 for $K = 16$ and $M = 11$), shows how the GMM and FV layers can efficiently reduce dimensionality. This is important in cases where the number of data samples is not very large, as is typically the case in the domain of mental illnesses.

4.2.5 Regression layers

In order to estimate the symptom severity in schizophrenia, we use two Fully Connected (FC) layers that receive as input the output of the FV layer. The first layer "FC1" is used for

estimating individual expression-related symptoms, while the second layer "FC2" estimates the total scale/symptoms score (e.g. CAINS Expression scale). The number of neurons in FC1 is adjusted according to the number of the estimated symptoms in each scale (flat affect, poor rapport, and lack of spontaneity and flow of conversation symptoms in the case of PANSS and 4 Expression symptoms in the case of CAINS). Two discriminative costs are used for training the regression layers as shown in Figure 4.1(b); the first for fine-tuning the GMM with the FV and FC1 layers in an end-to-end fashion, and the second for training the FC2 layer. The mean square error is used as the discriminative cost function, and is calculated as follows:

$$C_d(p,t) = \frac{1}{V}\sum_{v=1}^{V}\frac{1}{W}\sum_{w=1}^{W}(p_{vw} - t_{vw})^2, \tag{4.11}$$

where V denotes the total number of videos/patients in our training set, W is the number of symptoms estimated, and p and t represent the model's estimated symptom and the ground-truth value, respectively. The activation function used in FC1 and FC2 is the Rectified Linear Unit function. As the symptoms of schizophrenia have integer-based scores, the final outputs are rounded to the nearest integer during testing.

Number of trainable parameters in SchiNet. The GMM layer consists of K Gaussian distributions, and each distribution k has 3 trainable parameters w_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, where w_k is a scalar, $\boldsymbol{\mu}_k$ is a M-dimensional vector, and $\boldsymbol{\Sigma}_k$ is a diagonal matrix containing M trainable parameters (M is the number of detected AUs). The first regression layer (FC1) has $Q \times W$ dimensional weight matrix and a W-dimensional bias vector, where Q is the dimensionality of the FV representation and W is the number of estimated symptoms. The second regression layer (FC2) has W- and 1- dimensional weight and bias vectors, respectively. The total number of trainable parameters in SchiNet is $K(2M+1) + W(K(2M+1)+1) + (W+1)$. As an example, a network with $K = 16$, $M = 11$ and $W = 4$, has only 1849 trainable parameters.

4.3 Experiments and results

In this section, we first measure the correlations between facial expressions and different symptoms of schizophrenia. Then, we report the performance of the proposed SchiNet in

estimating symptom severity and compare it to other works in the literature.

4.3.1 Statistical analysis

The goal in this section is to calculate and examine how well a very simple feature extracted for each of the automatically detected AUs, namely, the frequency of the occurrence of the AU in question, correlate with the different symptom scales of schizophrenia. We show that for several symptoms, high and significant correlations with AUs are observed.

Symptom scales. In the NESS dataset that we use, the severity of symptoms of schizophrenia is assessed by two observer-rated scales, PANSS [75], and CAINS [61]. PANSS consists of a total of 30 symptoms divided into 3 scales: Negative (NEG), Positive (POS) and General Psychopathology (GEN). Out of the 30 symptoms, 7 are grouped to form the NEG scale, 7 form the POS scale, and the remaining 16 symptoms form the GEN scale. Each symptom in the PANSS is rated between 1 (absent) and 7 (extreme). On the other hand, CAINS consists of 13 symptoms, divided into 2 scales: Motivation and Pleasure (MAP), and Expression (EXP). MAP has 9 symptoms and EXP has 4 symptoms. Each symptom in the CAINS has a value between 0 and 4 (0=no impairment and 4=severe impairment). In PANSS, the total NEG, POS and GEN scores are the summation of the scores of the NEG, POS, and GEN symptoms, respectively. In CAINS, the total EXP score is the summation of the scores of the 4 EXP symptoms. More details on the PANSS and CAINS scales can be found in Appendix A.

Calculating correlations. We use the Spearmans's correlation for measuring the association between the ground-truth symptom levels and the activation frequency of each AU. In order to calculate the frequency, first we get a binary vector for each video frame, representing the presence or absence of each of the M expressions, and then compute the activation frequency as follows:

$$f_i = \frac{N_i}{N_{total}}, \quad i \in \{1, 2, ..., M\}, \tag{4.12}$$

where N_i is the number of frames for which AU i is activated and N_{total} is the total number

of video frames with a successful face detection. Note that the number of detected AUs M is 18 for the first AUs detection architecture (in Section 3.1), and 11 for the second architecture (in Section 3.2).

The faces of some patients are occluded by a wearable item (e.g. thick eyeglasses) – this sometimes results in the related AU being wrongly detected. In order to avoid these false detections, only frequencies that fall in the range of $-1.5\sigma_i \leq f_i \leq 1.5\sigma_i$ are considered, where σ_i is the standard deviation over the frequencies of AU i in the NESS dataset. Note that this step is applied only during statistical analysis and is replaced by the normalization step during symptom estimation.

Results. In our analysis, we measure all the possible correlations between the PANSS/CAINS symptoms and the detected AUs – however, we report only the significant correlations found, and discard the weak and insignificant ones, in order to have compact and focused tables of results. Table 4.3 and 4.4 show the significant correlations found between some CAINS and PANSS symptoms on the one hand, and AUs detected by the method in Section 3.1 (the one trained in controlled settings) on the other hand. In CAINS (Table 4.3), significant associations are found between mouth opening AUs (AU25 and AU26), which are commonly activated during patients' speech, and symptoms like quantity of speech and vocal expression. Also, a significant correlation is found between AU26 and the facial expression symptom. Furthermore, significant correlations are found between AUs and the total score of the CAINS and PANSS scales. For instance, brow lowerer (AU4) has significant associations with the CAINS-EXP and PANSS-GEN total scores.

Table 4.5 and 4.6 show the significant correlations found between AUs detected by the method in Section 3.2 (the one trained in the wild) on the one hand, and some symptoms in both of the CAINS and PANSS scales on the other hand. In CAINS (Table 4.5), significant associations are found between lips part (AU25), commonly activated during speech, and symptoms like quantity of speech, vocal expression, and facial expression. Similarly, in PANSS (Table 4.6), higher levels of symptoms like lack of spontaneity and flow of con-

Table 4.3: Correlations found between the **CAINS** symptoms and AUs detected by the method trained in **controlled** settings.

Symptoms \ Facial Expressions	AU4 – Brow Lowerer	AU15 – Lip Corner Depressor	AU25 – Lips Part	AU26 – Jaw Drop
EXP - Facial Expression	-	-	-	-0.34^*
EXP - Vocal Expression	-0.36^*	-	-	-0.32^*
EXP - Quantity of Speech	-	-	-0.33^*	-
MAP - Motivation for Close Family Relationships	-	-0.34^*	-	-
EXP - Total Score	-0.35^*	-	-	-

* indicates $p \leqslant 0.01$

Table 4.4: Correlations found between the **PANSS** symptoms and AUs detected by the method trained in **controlled** settings.

Symptoms \ Facial Expressions	AU4 – Brow Lowerer	AU26 – Jaw Drop	AU45 – Blink
GEN - Motor Retardation	-	-	-0.34^*
GEN - Unusual thought content	-	0.35^*	-
GEN - Total Score	0.34^*	-	-

* indicates $p \leqslant 0.01$

versation, poor rapport, and flat affect are associated with lower frequencies of the lips part. Moreover, symptoms related to the impairment in social interaction (e.g. poor rapport, flat affect, facial expression) are found to be correlated to smile and smile-related behaviour (cheek raiser). Finally, correlations are also found between AUs and the total score of the CAINS and PANSS scales. For instance, CAINS-EXP scale has significant associations with many AUs e.g. neutral expression, cheek raiser and lips part.

The two AUs detection architectures have different preprocessing steps, resulting in different number of patients being processed (74 for the first architecture and 91 for the second) – subsequently, the correlations obtained by the two architectures are not directly comparable. Moreover, the two architectures perform differently in AUs detection, that is, the first method can detect AUs at different levels of intensity, but it is working mainly on frontal or near-frontal

Table 4.5: Correlations found between the **CAINS** symptoms and AUs detected by the method trained **in the wild**.

Symptoms \ Facial Expressions	AU0 – Neutral Expression	AU6 – Cheek Raiser	AU7 – Lid Tightener	AU12 – Lip Corner Puller	AU25 – Lips Part	Smiling	AU43 – Eyes Closed
EXP - Facial Expression	$\mathbf{0.45}^{**}$	$\mathbf{-0.43}^{**}$	-	$\mathbf{-0.4}^{**}$	-0.33^{*}	$\mathbf{-0.42}^{**}$	-
EXP - Vocal Expression	$\mathbf{0.35}^{**}$	$\mathbf{-0.38}^{**}$	-0.34^{*}	-	$\mathbf{-0.41}^{**}$	-	-
EXP - Expressive Gestures	-	-0.32^{*}	-	-	$\mathbf{-0.43}^{**}$	-	-
EXP - Quantity of Speech	$\mathbf{0.38}^{**}$	-	-	-	$\mathbf{-0.41}^{**}$	-	-
MAP - Motivation for Recreational Activities	-	-	-	-	-	-	$\mathbf{-0.47}^{**}$
MAP - Frequency of Pleasurable Recreational Activities - Past Week	-	-	-	-	-	-	$\mathbf{-0.35}^{**}$
EXP - Total Score	$\mathbf{0.42}^{**}$	$\mathbf{-0.41}^{**}$	-	-	$\mathbf{-0.46}^{**}$	-0.29^{*}	-

** indicates $p \leqslant 0.001$, * indicates $p \leqslant 0.01$

Table 4.6: Correlations found between the **PANSS** symptoms and AUs detected by the method trained **in the wild**.

Symptoms \ Facial Expressions	AU0 – Neutral Expression	AU1 – Inner Brow Raiser	AU2 – Outer Brow Raiser	AU6 – Cheek Raiser	AU7 – Lid Tightener	AU25 – Lips Part	Smiling
NEG - Flat Affect	0.28^{*}	-	-	$\mathbf{-0.33}^{**}$	-	$\mathbf{-0.37}^{**}$	-0.29^{*}
NEG - Poor Rapport	-	-	-	$\mathbf{-0.36}^{**}$	-	-0.34^{*}	-0.28^{*}
NEG - Lack of Spontaneity and Flow of Conversation	0.32^{*}	-	-	-	-	-0.31^{*}	-
POS - Suspiciousness/Persecution	-	-	$\mathbf{0.36}^{**}$	-	-	-	-
GEN - Somatic Concern	-	0.29^{*}	-	-	0.33^{*}	-	-
GEN - Anxiety	-	-	0.29^{*}	-	-	-	-
NEG - Total Score	-	-	-	-	-	-0.30^{*}	-0.30^{*}
POS - Total Score	-	0.31^{*}	0.30^{*}	-	-	0.29^{*}	-
GEN - Total Score	-	-	-	-	$\mathbf{0.37}^{**}$	-	-

** indicates $p \leqslant 0.001$, * indicates $p \leqslant 0.01$

faces and in controlled settings, while the second method detects only moderate/intense AUs, but at different head poses and recording conditions – this affects the correlation values obtained by the two architectures. However, we still can see similarities in the correlations, e.g. the correlations found between the mouth opening AUs (AU25 and AU26), and some CAINS symptoms like quantity of speech, vocal expression, and facial expression. In general, facial expression analysis in the wild leads to more patients being analyzed in the NESS dataset, and more significant correlations compared to the analysis in controlled settings. In the next

section, we will show how each architecture affects the performance of the symptom severity estimation in schizophrenia.

We compare the correlations found in Table 4.5 and 4.6 to the ones reported in the literature, and specifically to the works that used automatic facial expression analysis in studying schizophrenia like [57, 135]. In [57], Hamm *et al.* compared patients with schizophrenia to healthy controls in terms of the temporal profiles of different AUs. [57] found that controls show more cheek raiser (AU6), lid tightener (AU7), and lip corner puller (AU12) than patients – this intersects with our findings, that is, we found that the frequencies of these AUs decrease with higher levels of symptoms like facial expression, vocal expression, or/and flat affect. Moreover, Tron *et al.* in [135] compared patients with severe negative symptoms to healthy controls in terms of the mean activity of several AUs. [135] found significant difference in smile activity between patients and controls (patients show less smiles). Similarly, we found that increased negative symptoms like flat affect, and PANSS-NEG and CAINS-EXP total scores are associated with decreases in the smiling behaviour. A similar finding was found for the lips part (AU25) behaviour (named lips up in [135]). On the other hand, [135] found significant difference in the frown (AU4) activity between patients and controls, while no association was found in our work. Also, our work found that patients with high severity show less cheek raiser (AU6) than those with low severity, while the opposite was found in [135] – these differences need further investigation in future studies.

The correlations found in Table 4.5 and 4.6 are also compared to those found in other works in psychiatry like [10, 21, 40, 133, 147]. In psychiatry, patients' non-verbal behaviour was manually annotated in terms of ECSI [132]. Annotated ECSI items are then grouped into 8 behaviour categories; affiliative, submission, prosocial, flight, assertion, gesture, displacement, and relaxation. Then, the correlations between these categories and the PANSS/CAINS subscales were measured. Each behaviour category includes facial expressions as well as head and body gestures. Here, we compare the categorical correlations found in psychiatry to the correlations of the category-related expressions found in our work. Although, the correlations are

not directly comparable, we still can see some similarities. For instance, in [10, 133, 147] the prosocial category was negatively associated with the CAINS-EXP/PANSS-NEG total score, and similarly in our work the smiling behaviour, which is one of the prosocial behaviour, was negatively associated with the CAINS-EXP and PANSS-NEG scores. Also, our work and [147] found no association between the assertion category or the frown (an assertion-related expression), and the CAINS/PANSS subscales. Furthermore, [10] found that the affiliative category is positively correlated to the positive symptoms, and similarly our work found that AU1 and AU2 (affiliative-related expressions) are positively correlated to the PANSS-POS total score.

On the other hand, we can see differences between the correlations found in our work and other works in psychiatry. For instance, [40, 147] found no correlation between the relaxation category and the CAINS/PANSS subscales, while in our work a significant correlation was found between the neutral expression (a relaxation-related expression) and the CAINS-EXP total score. A possible reason is that [40, 147] used a scoring system that annotates behaviour present more than once in a segment of 30 seconds with the same score – this cannot capture the variance in activity for frequent behaviour like neutral expression. Moreover, [21, 40, 147] found significant correlation between the flight category and the CAINS-EXP and PANSS-NEG total scores, however, no correlation was found in our work between the closure of eyes (a flight-related expression) and the CAINS-EXP/PANSS-NEG subscales – these differences in correlations need further investigation.

We have explored in our work how some AUs (that are not part of ECSI) like AU5, AU6, AU7 and AU25, correlate to symptoms of schizophrenia. AU6, AU7 and AU25 showed significant correlations with many symptoms in the PANSS and CAINS scales, while AU5 showed no correlations. Hence, we encourage researchers to include AU6, AU7 and AU25 in their behaviour analysis.

4.3.2 Symptom severity estimation

Among the different types of symptoms of schizophrenia, negative symptoms are particularly difficult to assess and quantify. The assessment requires the quantification of observed verbal and especially non-verbal behaviour so that ratings commonly involve a large degree of subjectivity. Thus, an objective method for assessing these symptoms would be an important achievement. It is being debated as to what extent negative symptoms do or do not change in treatment interventions [112], and measures that are obtained in an automatic way may establish symptoms with higher accuracy and reliability and therefore help to clarify whether changes do or do not occur. Motivated by that, we focus in this chapter on assessing the highly correlated negative symptoms in both the CAINS and PANSS interviews through the automatic analysis of the video interviews.

Training settings. For CAINS, the GMM and FV layers are trained firstly end-to-end with the FC1 layer for estimating the 4 EXP symptoms. Then, the GMM and FC1 parameters are kept fixed and the FC2 layer is trained on estimating the total EXP score. Similarly, the three highly correlated NEG symptoms in PANSS, namely flat affect, poor rapport, and lack of spontaneity and flow of conversation, are estimated at the FC1 layer, and the total NEG score is estimated at the FC2 layer. Note that the number of neurons in FC1 layer is equal to the estimated symptoms at each scale.

We use the Theano/Lasagne framework [127, 39] for implementing the GMM, FV, and FC layers. The number of GMM components (K) is set to 16. Following [111], we use variance flooring to avoid instability in the calculations – the minimum variance allowed is 0.001. Moreover, whenever the posterior is below a threshold of 10^{-4} it is set to zero – this leads to a sparser FV. The GMM-FV-FC1 layers are trained using Stochastic Gradient Descent (SGD) with momentum $m = 0.9$ and learning rate $lr = 0.005$ for CAINS and 0.001 for PANSS. The FC2 layer is trained also using SGD with $m = 0.9$ and $lr = 0.01$. Finally, in the case of the CAINS scale, a scaling factor is learned for the training set and applied at testing, so as to scale the output values in the range between the minimum (0) and the maximum (4) values.

Leave-one-subject-out is used for validating and testing our architecture.

Performance measures. Three measures are used for reporting the performance of the symptom severity estimation using as ground truth the psychiatrists' assessments. Following [135, 137], we use the Pearson's Correlation Coefficient (PCC) and, in addition to it, we report the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). The former (i.e. the MAE) is less sensitive to outliers, while the latter (i.e. the RMSE) emphasizes more on larger differences.

Effect of the AUs detection architecture. In this experiment we evaluate how the two proposed architectures for AUs detection (described in Chapter 3) perform on symptom severity estimation. For the first method (in Section 3.1), we use the "Full" architecture consisting of 8 neural networks for analyzing facial expressions of the patients. This method uses two output neurons to predict, respectively, the presence and absence of each expression/AU – during testing the one with the highest probability is selected. Here, in order to get an expression probability between 0-1, the presence-probability is divided by the sum of both the presence- and absence-probabilities. 18 AUs are detected by the first method, while 11 by the second method. For both methods, the mean over each AU is subtracted (normalization step), and then the AUs probabilities are used as input to the GMM layer.

As the number of patients and AUs analysed by the two architectures vary, only the ones that are common in both methods are used in the comparison. Based on that, 69 patients and 8 common AUs (shown in Table 3.9) are selected for symptom estimation. The number of GMM components is set to 12 in this case, as fewer patients and AUs are analysed. Table 4.7 and 4.8 show the estimation results of the CAINS and PANSS symptoms, respectively, using both architectures, the one working in controlled settings (i.e. the first method) and the other working in the wild (i.e. the second method). From the comparison, we can see that the method trained in the wild leads to better symptom estimation in all the estimated symptoms. This illustrates the positive impact of training the AUs detection architecture using data captured in the wild. In addition, the second method leads to more significant correlations

Table 4.7: Comparison between the two proposed AUs detection methods on estimating **CAINS-EXP** symptoms.

	First method (Full)			**Second method**		
	PCC	MAE	RMSE	PCC	MAE	RMSE
EXP - Facial Expression	0.37	0.80	1.07	**0.42**	**0.74**	**0.99**
EXP - Vocal Expression	0.25	0.93	1.22	**0.30**	**0.75**	**1.13**
EXP - Expressive Gestures	0.04	1.07	1.37	**0.34**	**0.99**	**1.19**
EXP - Quantity of Speech	**0.42**	1.07	1.37	0.39	**0.91**	**1.22**
EXP - Total Score	0.29	3.06	3.80	**0.42**	**2.90**	**3.61**

Table 4.8: Comparison between the two proposed AUs detection methods on estimating **PANSS-NEG** symptoms.

	First method (Full)			**Second method**		
	PCC	MAE	RMSE	PCC	MAE	RMSE
NEG - Flat Affect	0.21	0.97	1.31	**0.32**	**0.96**	**1.27**
NEG - Poor Rapport	0.25	0.91	1.22	**0.41**	**0.75**	**1.13**
NEG - Lack of Spontaneity and Flow of Conversation	0.13	**1.28**	1.60	**0.24**	**1.28**	**1.54**
NEG - Total Score	0.08	4.00	4.88	**0.40**	**3.35**	**4.27**

and can analyze more patients (91 patients vs 74 patients), compared to the first method.

Comparison to state of the art. In this section we compare the proposed SchiNet with two other methods that have been proposed in the literature for symptom severity estimation, namely, Tron *et al.* [135, 137]. We have re-implemented both methods, and for a fair comparison, the pre- and post-processing steps (e.g. normalization, scaling) applied in the SchiNet, are also applied to them. In [137], Tron *et al.* used the "Elbow criterion" for selecting the best number of clusters – here, we tried different number of clusters in the range of 2-24, and report the best results (obtained for 12 clusters). Furthermore, since the methods in [135, 137] estimate specific symptoms of schizophrenia and not the total score, we discard from the comparison the total CAINS-EXP and PANSS-NEG scores. In this comparison we use the AUs probabilities of the 91 patients, that are analyzed by the AUs detection architecture trained in the wild. Table 4.9 and 4.10 summarize the results. SchiNet outperforms the other methods in the 3 PANSS-NEG symptoms, and in 3 out of the 4 CAINS-EXP symptoms. The extracted statistical features using the GMM and FV layers show better performance compared to the

Table 4.9: Performance of the SchiNet as well as other state-of-the-art methods on the **CAINS-EXP** symptoms.

	Tron *et al.* [135]			**Tron *et al.* [137]**			**SchiNet**		
	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE
EXP - Facial Expression	0.37	0.80	1.03	0.36	0.75	1.07	**0.46**	**0.66**	**0.93**
EXP - Vocal Expression	0.23	0.87	1.23	0.26	0.86	1.22	**0.27**	**0.77**	**1.10**
EXP - Expressive Gestures	0.36	**0.85**	1.19	**0.38**	0.91	1.22	0.36	0.90	**1.15**
EXP - Quantity of Speech	0.27	1.09	1.43	0.25	1.02	1.36	**0.30**	**0.98**	**1.30**
EXP - Total Score	-	-	-	-	-	-	**0.45**	**2.67**	**3.34**

Table 4.10: Performance of the SchiNet as well as other state-of-the-art methods on the **PANSS-NEG** symptoms.

	Tron *et al.* [135]			**Tron *et al.* [137]**			**SchiNet**		
	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE
NEG - Flat Affect	0.37	0.90	1.28	0.11	0.99	1.36	**0.42**	**0.84**	**1.18**
NEG - Poor Rapport	0.20	0.98	1.31	0.15	1.01	1.26	**0.27**	**0.85**	**1.20**
NEG - Lack of Spontaneity and Flow of Conversation	0.13	1.37	1.69	0.09	1.32	1.62	**0.25**	**1.25**	**1.51**
NEG - Total Score	-	-	-	-	-	-	**0.29**	**3.30**	**4.17**

hand-crafted features.

For estimating the total PANSS-NEG score, the proposed architecture uses the 3 out of 7 PANSS-NEG symptoms that are highly correlated with AUs, as input to the FC2 layer. In Table 4.11, we report the estimation results in two additional settings. First, by estimating directly the total score from the FV representation, that is by using a single fully-connected layer (FC1) with a single output. Second by estimating the total score from all the PANSS-NEG symptoms. In this latter setting, we first train the SchiNet on estimating the 7 PANSS-NEG symptoms at FC1 layer, and then estimating the total score at FC2 layer. In both cases, the results were worse. In the first case a possible reason is that NEG symptoms have more significant correlations with AUs than the total NEG score, so estimating symptoms first, helps a lot in estimating the total score. In the second case a possible reason is that 4 out of the 7 NEG symptoms are not correlated to AUs, making the training of the FC2 layer worse.

Table 4.11: The severity estimation results of the total **PANSS-NEG** score, obtained by the SchiNet in different settings.

Input	SchiNet		
	PCC	MAE	RMSE
FV representation	0.08	3.35	4.37
FC1 layer with 3 symptoms	**0.29**	**3.30**	**4.17**
FC1 layer with 7 symptoms	0.18	3.37	4.37

4.4 Conclusion

Our work in this chapter aims to develop an architecture that is capable of using quantified patient behaviour for estimating the severity of different symptoms. To this end, interviews of symptom assessment recorded at different places in the UK were used in our analysis, in conditions that are similar to real clinical settings. Two different architectures are used for detecting patients' facial expressions. Then, the detected expressions are used as input to a neural network, that extracts compact statistical features and estimates symptoms of schizophrenia. We estimate expression-related negative symptoms in two different assessment interviews, PANSS and CAINS.

Our experimental results show many findings. First, we show that the proposed method for AUs detection in the wild performs better on symptom severity estimation than other method that was trained using data captured in a controlled environment. This underlines the importance of training with data collected in the wild. Second, significant correlations are found between symptoms and the frequency of occurrence of automatically detected facial expressions/AUs – this confirms that symptom levels of patients with schizophrenia are expressed in the degree of their impairments in expression of emotion and social interaction. Third, several symptoms in the PANSS and CAINS interviews can be estimated with a MAE less than 1 level. All of that leads to a conclusion that quantified patient behaviour with a well-trained deep architecture can be used for estimating negative symptoms of schizophrenia – the latter is a challenging task in clinical settings – and may be used as an objective method to establish changes during treatment.

Although our architecture shows promising results in symptom estimation, comparing the correlations between the automatic estimations and professional assessment (reaching at most to 0.46), to the correlations between assessments of different professionals that have annotated the NESS dataset [106] (equals to 0.85), shows that automatic estimation of symptom severity needs further improvement to reach human level performance. In order to improve the performance of symptom severity estimation, we suggest for future work improving the performance of the AUs detection method, by moving from static to temporal analysis in the wild. In addition, extending the behaviour analysis to include body gestures and vocal expressions (besides facial expressions).

CHAPTER 5

Treatment outcome estimation in schizophrenia

Contents

Negative symptoms of schizophrenia include expressive deficits that are marked by a reduction in patients' non-verbal behaviour. Analysing automatically non verbal behaviour, and in particular facial expressions, and exploiting the results for classifying (patients vs non-patients) or/and estimating symptom severity has shown promising results in Chapter 4, as well as in other works in the literature [135, 137]. The proposed methods for symptom estimation could be used for monitoring the changes in patient's symptom level during treatment interventions (i.e. the treatment outcome), by estimating the symptom level before and after

Parts of this chapter have been published in [19] and [20].

treatment, and then comparing the estimated levels. However, they do not perform well because the change in these symptoms is typically small and falls within their margin of error [112].

In this chapter we propose two Deep Learning architectures for addressing directly the problem of treatment outcome estimation in schizophrenia – more specifically, the proposed methods analyze jointly two videos of the same patient, one before and one after the treatment, and gives as output the treatment outcome, that is a binary label that encodes whether a symptom has improved or not. In both architectures, the patient's facial expressions in both videos are first detected, and then used as input to a deep neural network. To the best of our knowledge, these are the first works to address directly the problem of treatment outcome estimation in schizophrenia. The two architectures can be summarized as follows:

- Our **first** proposed architecture uses stacked Recurrent Neural Networks for learning local and global differences in patient's behaviour (facial expressions) before and after treatment. Specifically, a Gated Recurrent Unit (GRU) is used for learning the local differences in the patient's behaviour over short concatenated clips/segments from both videos. Then, another GRU uses the clip-level features for learning global (i.e. patient-level) features, and outputs the treatment outcome. This architecture is called "Stacked-RNNs". Stacked-RNNs assumes that the patient's videos are aligned and have equal length (videos with different lengths are clipped).

- The **second** architecture, named Temporal Attentive Relation Network (TARN), learns to align and compare representations (i.e. videos) of variable temporal length. The architecture consists of two modules: the embedding module and the relation module. In the embedding module, a GRU is used to extract short representations/embeddings over the facial expressions detected in short clips/segments of videos. In the relation module, a segment-by-segment attention mechanism is used first to align segment embeddings from the pair of videos (recorded before and after treatment). Then, the aligned segments are compared. The effect of using different comparator functions is explored.

Finally, the comparison outputs are aggregated using a deep neural network consisting of two fully-connected layers and an average pooling layer – this network gives as output the treatment outcome.

The two architectures have two main differences. First, Stacked-RNNs assumes that the patient's interviews are aligned and have equal lengths – this requires clipping videos of different lengths, and losing by that possibly useful information. On the other hand, TARN uses an attention mechanism for aligning and comparing videos of different lengths, avoiding by that any information loss. Second, TARN is trained in an end-to-end fashion, while Stacked-RNNs is trained in two steps, and consequently TARN is easier to train and test.

In the two architectures, videos are compared segment-wise for two main reasons. First, decomposing the video comparison problem into several subproblems (i.e. segment-to-segment comparison) showed good performance in [20], as this can be considered a way for augmenting the data without including additional samples – which helps in improving the training process and reducing overfitting. Second, encoding the video using segment-wise representations benefit the video-to-video comparison [20], compared to using a single video-wise representation – as video-wise representations can not capture the fine-grained information that exists in the video.

The architectures are trained in a patient-independent manner on a dataset of 88 patients with 176 video interviews – two interviews for each patient, one before and one after completing a 10-week period of treatment. The videos were recorded in uncontrolled conditions and in settings that are similar to real clinical ones. We estimate the treatment outcome of negative symptoms from two symptom assessment interviews; Clinical Assessment Interview for Negative Symptoms (CAINS) [61], and Positive and Negative Syndrome Scale (PANSS) [75]. Experimental results show that the proposed architectures achieve promising results for treatment outcome estimation over the different symptoms.

The rest of this chapter is organized as follows: we describe the first proposed architecture

(Stacked-RNNs) for treatment outcome estimation in Section 5.1. In Section 5.2, we present the second proposed architecture (TARN). In Section 5.3, we report the experimental results for both architectures. Finally, in Section 5.4 we conclude the chapter.

5.1 Stacked RNNs for treatment outcome estimation

In this section we present the first proposed architecture for treatment outcome estimation. Figure 5.1 shows an overview of the architecture. The architecture takes as input 2 video interviews of a patient, one recorded before the treatment (video-1) and the other recorded after (video-2), and outputs the treatment outcome, that is, either improved (i.e. symptom level went down) or not improved (i.e. symptom level stayed the same or went up). That is, it directly addresses the problem of treatment outcome estimation, posing it as a binary classification problem. The architecture consists of 4 stages; preprocessing, automatic facial expression analysis, feature selection, and sequence learning using Recurrent Neural Networks.

5.1.1 Preprocessing Steps

We first slice the videos into fixed length clips of 15 seconds each. The number of clips is kept fixed in the videos of the same patient. To deal with a pair of videos with different lengths, we divide the short video into N clips without overlap or spacing between the clips, and the long video into N equally-spaced clips, as shown in Figure 5.1. The sliced clips are then down-sampled by a factor of 3 to reduce the processing time. No difference in performance is noticed with the down-sampling in our initial experiments. A pair of clips (one from each video) is then passed to the next processing steps. Note that the number of clips N varies across the different patients.

For each frame in the paired clips, we detect the patient's body using [88]. We then extend the bounding-box of the detected body by a factor of 1.2 to ensure that the whole head is included, and then within the body region we apply the preprocessing steps of the used architecture for facial expression analysis (including face detection and scaling).

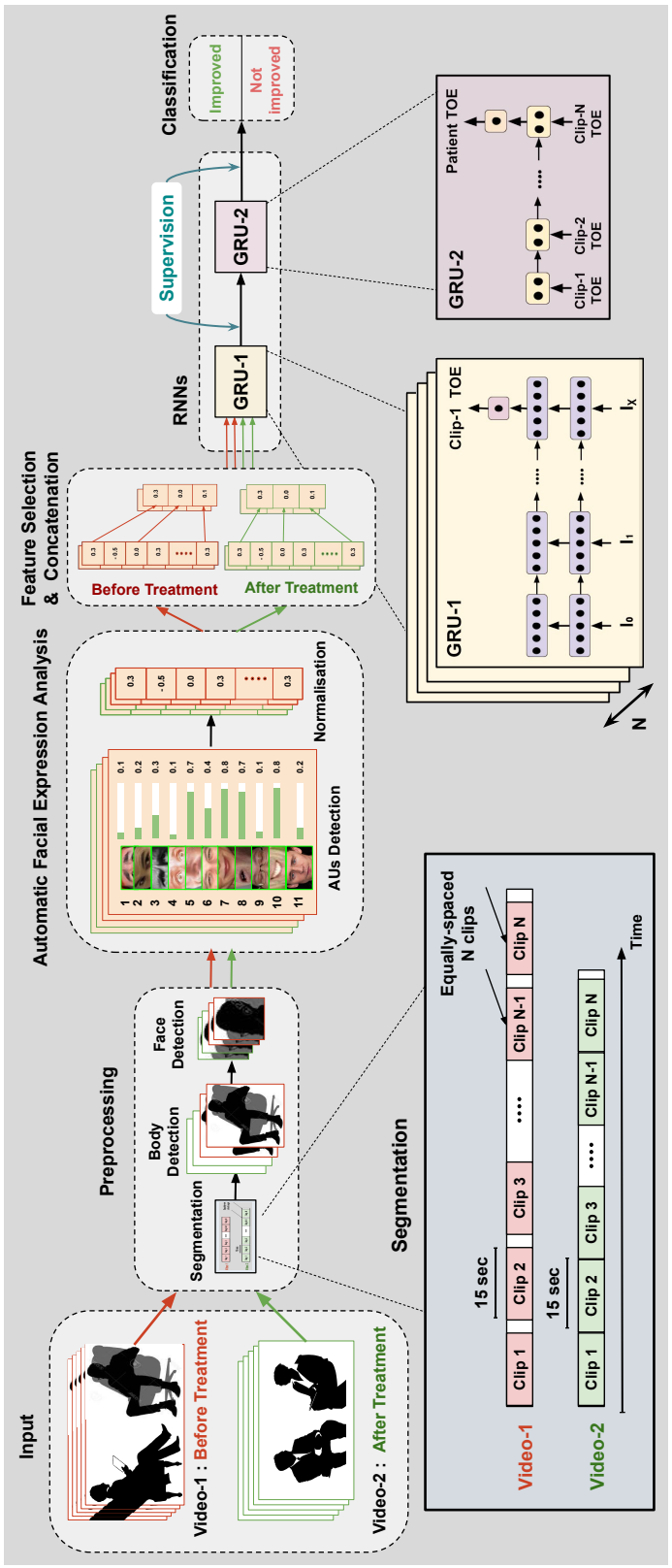


Figure 5.1: The proposed Stacked-RNNs architecture for treatment outcome estimation in schizophrenia.

In some videos, the camera is positioned in such a way that makes the patient's face hard to be detected. In those cases, even if the face is detected in some frames, it is hard to be further analysed in terms of facial expressions. Therefore, we consider only the videos in which we can successfully detect faces in more than 90% of the frames – this leads to 74 patients out of 88 included in our analysis. Note that video-1 and video-2 of each patient should meet this condition in order for the patient to be included.

5.1.2 Automatic facial expression analysis

In the second stage of our proposed architecture, we use automatic facial expression analysis for extracting low-level features from the paired clips. Specifically, we use the AUs detection architecture trained in the wild (explained in Section 3.2) for analyzing the patients' facial expressions. For each frame in the clips, we get an 11-dimensional feature vector, 10 dimensions corresponding to the probabilities of the presence of 10 AUs and one corresponding to the probability of the presence of a smile [68]. Note that in this chapter we only use the AUs detection architecture trained in the wild as it has shown better performance on symptom severity estimation compared to the other architecture trained in controlled settings, in Chapter 4.

Finally, in order to reduce the effect of camera viewpoints, illuminations levels, or/and occlusions by wearable items (e.g. sunglasses), for each video, the mean over each AU is calculated and subtracted from the AUs probabilities in the whole video (normalisation step). The normalised probabilities from video-1 and video-2 are concatenated at each time step and used as input to the next stage.

5.1.3 Feature selection

As the dimensionality of the features (AUs) is doubled by concatenation, and as a relatively small number of patients are available for training – feature selection is a crucial processing step in our architecture. Sequential forward feature selection is used for selecting the most relevant AUs to the estimated symptoms. The criterion for selecting features is the classification

performance on the validation set. The selected AUs from video-1 and video-2 are concatenated and used as input to the Recurrent Neural Networks (RNNs). Note that the same AUs are selected in both videos.

5.1.4 Sequence learning using stacked RNNs

RNNs is a class of Deep Neural Networks that is used for learning sequential information. Two popular models of RNNs are Long Short-Term Memory (LSTM) [60] and Gated Recurrent Unit (GRU) [26]. These models can learn long temporal dependencies without having the vanishing and the exploding gradient problem, through using gates with learnable parameters to control the information flow between time steps. In our architecture, we adopt a GRU to learn the temporal dynamics of the patients' facial expressions, as it has fewer parameters and generalises better on small datasets, in comparison to LSTM.

Our architecture consists of two stacked GRUs (GRU-1, GRU-2), shown in Figure 5.1, and takes as input pairs of sequences of facial expressions and outputs a soft decision, corresponding to whether the facial expressions in the second sequence (video-2) indicate an improvement in the symptoms in comparison to the first (video-1). That is, it treats the treatment outcome estimation as a binary classification problem using RNNs.

The first network (GRU-1) is used as a local (clip-level) feature extractor in our architecture. More specifically, GRU-1 is trained using the selected AUs probabilities in the pairs of clips for clip-level treatment outcome estimation. During training, GRU-1 is supervised by the patient treatment outcome (improved or not-improved). GRU-1 consists of a GRU layer with 16 hidden units, and a fully-connected layer with a single sigmoid unit for classification.

The second network (GRU-2) is used as a global feature extractor. In particular, GRU-2 uses clip-level features/estimations for learning global (i.e. patient-level) features, and outputs a soft binary label corresponding to the treatment outcome. GRU-2 consists of a GRU layer with 2 hidden units, and a sigmoid classification layer.

5.2 TARN: Temporal attentive relation network for treatment outcome estimation

In this section we introduce a novel deep architecture, named Temporal Attentive Relation Network (TARN) for the problem of treatment outcome estimation. Figure 5.2 shows an overview of the network. TARN takes as input two video interviews recorded for a patient before and after treatment (video-1, video-2), and gives as output a binary score representing if the patient got improved by the given treatment or not. TARN learns to align and compare representations (i.e. two videos) of variable temporal length. TARN consists of two modules: the embedding module and the relation module. These modules are explained in detail in the following subsections.

5.2.1 Embedding module

The patients' videos are sliced into fixed length segments of 15 seconds each, with no overlap or spacing between the segments, as shown in Figure 5.2. In this case, videos of the same patient might have different number of segments. The rest of the preprocessing steps remains the same as the ones explained in Section 5.1.1. The video segments are then fed to the AUs detection architecture trained in the wild. For each frame, we get 11-dimensional feature vector representing the probabilities of 11 AUs. The AUs detection architecture acts as a low-level feature extractor in the embedding module.

Then, a uni-directional GRU of size 4 uses the low-level features for learning high-level ones. Specifically, the GRU summarizes the AUs probabilities in the short video segments, and gives as output at the last time step of each segment, a short representation/embedding.

5.2.2 Relation module

In order to compare video-1 and video-2, we first align their segments by using a segment-by-segment attention layer. The attention layer maps the video-2 embeddings to have the same number of segment-embeddings as video-1. Second, each segment in video-1 is compared to

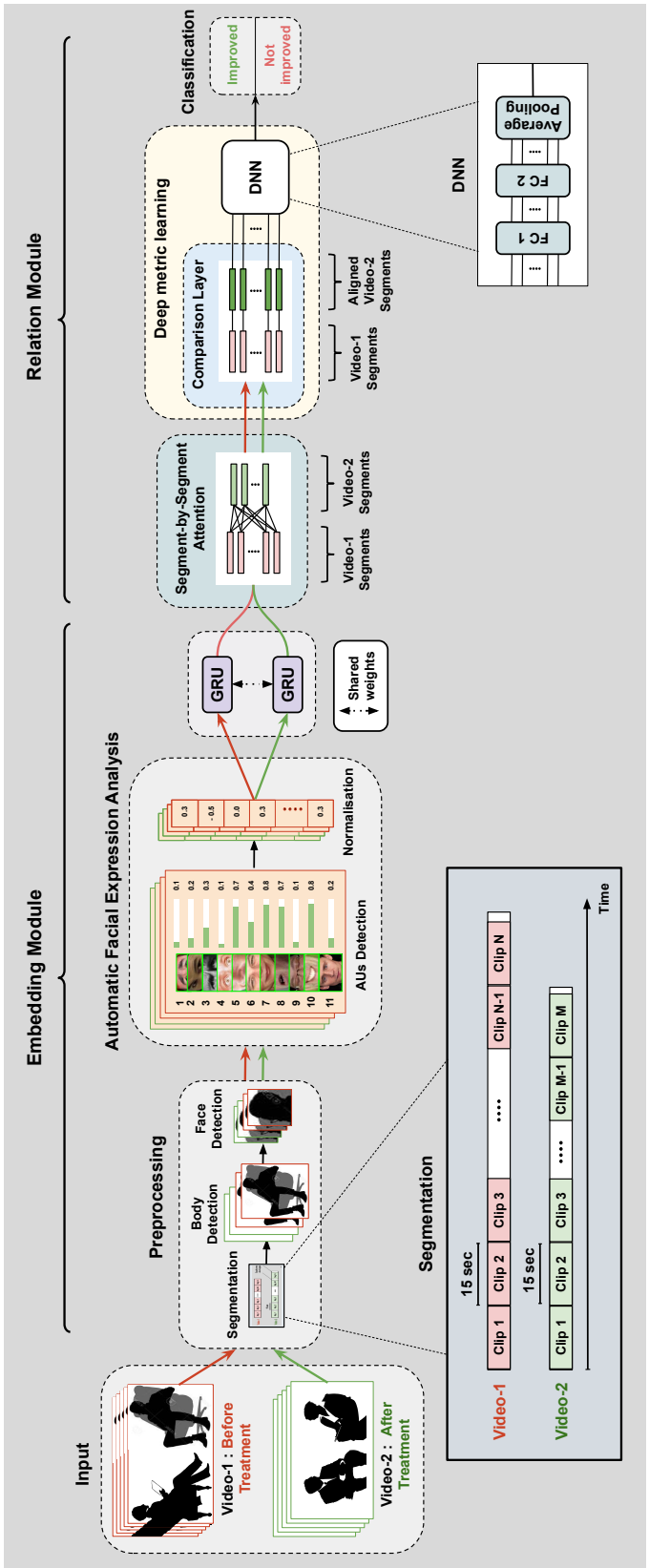


Figure 5.2: The proposed TARN architecture for treatment outcome estimation in schizophrenia.

the corresponding aligned segment in video-2. Third, the comparison outputs are aggregated by a deep neural network – this network learns a deep metric for video comparison, and gives at its output a relation score representing the treatment outcome.

Segment-by-segment attention. Several recent works in text sequence matching and textual entailment used an attention mechanism, named word-by-word attention, to align the words of two given sentences [12, 104, 109, 144]. Similarly, as shown in the corresponding block of Figure 5.2, we adopt the word-by-word attention in our architecture to align the video-1 and video-2 segment-embeddings (i.e. segment-by-segment attention). Let us consider a video $\boldsymbol{S} \in \mathbb{R}^{N \times d}$ recorded before treatment, and a video $\boldsymbol{Q} \in \mathbb{R}^{M \times d}$ recorded after treatment, where each row in $\boldsymbol{S}$ and $\boldsymbol{Q}$ represents a segment-embedding vector of dimension d, and where N and M denote the number of segments in videos $\boldsymbol{S}$ and $\boldsymbol{Q}$ respectively. The segment-by-segment attention is calculated as follows:

$$\boldsymbol{A} = \text{softmax}((\boldsymbol{S}\boldsymbol{W} + \boldsymbol{b} \otimes \boldsymbol{e}_N)\boldsymbol{Q}^T), \tag{5.1}$$

$$\boldsymbol{H} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{S}, \tag{5.2}$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ are parameters to be learned, and the operator "$\otimes \boldsymbol{e}_N$" repeats the bias vector $\boldsymbol{b}$, N times to form a matrix of dimension $N \times d$. $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ is the attention weight matrix and $\boldsymbol{H} \in \mathbb{R}^{M \times d}$ is the aligned version of $\boldsymbol{S}$. Each row vector in $\boldsymbol{H}$ is a weighted sum of the $\boldsymbol{S}$ segment-embeddings, and represents the parts of $\boldsymbol{S}$ that are most similar to the corresponding row vector (segment-embedding) of $\boldsymbol{Q}$. The row vectors of $\boldsymbol{Q}$ and $\boldsymbol{H}$ are used as inputs to a comparison layer.

Deep metric learning. The relation module performs deep metric learning by using a comparison layer and a non-linear classifier on the top of it. The comparison layer calculates a distance/similarity measure between each of the M segments (row vectors) of $\boldsymbol{Q} \in \mathbb{R}^{M \times d}$ and $\boldsymbol{H} \in \mathbb{R}^{M \times d}$. This measure, as described in [144], can be based on one of the following operations: multiplication (Mult), subtraction (Subt), neural network (NN), subtraction and multiplication followed by a neural network (SubMultNN), or Euclidean distance and cosine

similarity (EucCos). In addition to the previous measures, we also explore the effect of using the simple concatenation (Conc) process, that is used in the Stacked-RNNs architecture. Since the measure is estimated at each of the M pairs of segments, the output of this layer has also M dimensions. This layer acts as an intermediate stage that produces low-level representations of the comparisons between the video-1 and video-2 segments.

Unlike other works that used a linear classifier or a fixed metric to compare different data samples [77, 117], we follow [124] and use a neural network for deep metric learning. That is, the outputs of the comparison layer are passed to the neural network for learning a global deep metric over the entire videos. We use two Fully-Connected (FC) layers and an average pooling layer for aggregating the comparison outputs. The pooling layer gives as output the final relation score, corresponding to whether the facial expressions in video-2 indicate an improvement in the symptoms in comparison to video-1. The two FC layers are of size four and one.

The TARN architecture (excluding the AUs detection architecture) is trained in an end-to-end fashion. The binary cross-entropy is used as the cost function, and is calculated as follows:

$$L(t, q) = -\frac{1}{B}\sum_{b=1}^{B}(t_b \log q_b + (1 - t_b)\log(1 - q_b)), \tag{5.3}$$

where B denotes the batch size, t the target treatment outcome, and q the predicted treatment outcome.

5.3 Experiments and results

5.3.1 The data of schizophrenia

In this work we use recordings and symptom annotations from the "NESS" trial [106], that was collected for evaluating body psychotherapy as a treatment for negative symptoms of schizophrenia. We choose the NESS trial for our analysis as it was recorded in realistic conditions and in settings that are similar to the ones found in clinics and hospitals. In total 275 parti-

cipants were included in the NESS trial. All participants were diagnosed with schizophrenia, and had a total negative symptoms score ≥ 18 on the PANSS scale. The participants were assessed 3 times during the NESS trial; before starting the treatment (baseline), after completing a 10-week treatment (end of treatment), and 6 months after the end of the treatment (6 months follow-up). Several scales were used for measuring the outcome of the treatment such as PANSS [75] including negative, positive and general psychopathology symptoms, and CAINS [61] including experience-related and expression symptoms. Researchers/psychologists conducted the assessment interviews in a structured way that is similar to real life clinical settings.

The participants were video-recorded during the PANSS and CAINS assessment. The NESS trial contains recordings for 110 patients at baseline, 93 patients at end of treatment, and 69 patients at 6 months follow-up – as not all of the patients accepted to be recorded at all sessions. In order to build a dataset for the problem of treatment outcome estimation, we select the patients who have been recorded at two out of the three sessions – this leads to a dataset of 88 patients, where each patient has two videos (commonly one at baseline and the other at the end of the treatment). Out of the 88, we considered only 74 patients for whom we can successfully detect faces in more than 90% of the frames of their videos. Most of the videos were recorded at a frame rate of 25 f/s and a resolution of 1920×1080. The average length of all the videos in our dataset is 42 minutes. More information about the NESS trial can be found in [106].

Each of the 74 patients has two symptom severity scores, one is given before treatment, while the other after treatment. The patient is considered improved by the given treatment if the symptom score after the treatment is less than the score before the treatment (i.e. the symptom level went down), and not improved if the symptom score after the treatment is more than or equal to the score before the treatment (i.e. the symptom level stayed the same or went up). Subsequently, the treatment outcome in our analysis is given a binary label, either "1" to represent symptom improvement or "0" to represent no symptom improvement. The distribution of the treatment outcome labels for the expression-related symptoms in the

Table 5.1: The distribution of the treatment outcome labels for the CAINS-EXP symptoms.

Labels / CAINS Symptoms	Not improved	Improved
EXP - Facial Expression	46	28
EXP - Vocal Expression	46	28
EXP - Expressive Gestures	50	24
EXP - Quantity of Speech	48	26

Table 5.2: The distribution of the treatment outcome labels for the expression-related PANSS-NEG symptoms.

Labels / PANSS Symptoms	Not improved	Improved
NEG - Flat Affect	45	29
NEG - Poor Rapport	47	27
NEG - Lack of Spontaneity and Flow of Conversation	47	27

PANSS and CAINS scales, is shown in Table 5.1 and 5.2, respectively.

5.3.2 Training settings

We use 74 pairs of video interviews (one for each of the 74 patients) for training and testing our architectures using a Leave-One-Subject-Out (LOSO) protocol. More specifically, for each fold in LOSO, 67 patients are used for training, 6 patients for validation, and 1 patient for testing. We augment the dataset with extra samples by considering each pair of videos in the training, validation and testing sets as two data samples. Specifically, we change the order of each pair of video-1 and video-2, and the ground truth label accordingly to get an extra data sample.

We train the proposed architectures for estimating the change in negative symptoms, especially symptoms annotated based on patients' non-verbal behaviour during symptom assessment interviews. In particular, 4 expression symptoms (i.e. the Expression scale) in the CAINS interview [61], and 3 symptoms (flat affect, poor rapport, lack of spontaneity and flow of conversation) in the PANSS interview [75], are estimated. Note that a separate network is trained for each symptom.

The **Stacked-RNNs** architecture is trained in two steps. First, pairs of sliced clips partitioned from all patients' videos are used for training GRU-1. The output of GRU-1, that is the clip-level estimations, are then used to train GRU-2. We use the binary cross-entropy classification cost for both networks. We train the GRUs using stochastic gradient descent with adaptive learning rate (RMSprop [129]), with a decay coefficient set to 0.7, and gradient clipping to 100. The initial learning rate is set to 0.005 for GRU-1, and 0.01 for GRU-2. The batch size is set to 256 sequences for GRU-1 and the training set size (i.e. 67×2=134 batches) for GRU-2.

The **TARN** architecture is trained in an end-to-end fashion. Adam optimizer [76] with an initial learning rate set to 0.01 and gradient clipping to 1 is used in the training. The batch size is set to half the training set size (i.e. 67 video pairs). Dropout [120] with a probability of 0.2 is used for regularization. Finally, we use the Theano/Lasagne framework [127, 39] for implementing the Stacked-RNNs architecture, while the PyTorch library for implementing the TARN architecture (as Theano recently become outdated).

5.3.3 Results

Performance Measures. We choose the accuracy and F1-score of both classes to evaluate the performance of the proposed architectures. Accuracy is a widely-used measure in classification problems. However, it could not reflect well the performance over the minority class when the data is highly imbalanced. In our case, the ratio of the negative to positive examples over the different symptoms is roughly 2:1. Hence, we report both the accuracy and F1-score. We use $F1_P$ to refer to the F1-score of the positive class, and $F1_N$ to the F1-score of the negative class.

TARN ablation studies. In our first experiment, we investigate the impact of the various functions that can be used as a distance/similarity measure in the TARN comparison layer, on the treatment outcome estimation performance. Specifically, we compare the five different distance measures (Mult, Subt, NN, SubMultNN, EucCos) mentioned in [144], in addition

Table 5.3: Performance of the TARN architecture when using different similarity/distance measures in the comparison layer.

Symptom	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			Average over all symptoms		
	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc
Conc	0.37	0.75	0.64	0.40	0.67	0.57	**0.43**	0.73	0.63	0.43	0.73	0.63	0.41	0.72	0.62
NN	0.46	0.71	0.62	0.29	**0.75**	**0.63**	0.36	0.75	0.64	**0.47**	0.62	0.56	0.40	0.71	0.61
Subt	0.42	0.73	0.63	0.33	0.64	0.53	0.37	0.70	0.59	0.42	0.78	0.68	0.39	0.71	0.61
Mult	0.39	**0.76**	**0.66**	0.29	0.73	0.61	0.42	0.76	0.66	0.41	0.78	0.68	0.38	**0.76**	0.65
SubMultNN	0.41	0.74	0.64	0.26	0.74	0.62	0.34	0.77	0.66	0.37	**0.80**	**0.70**	0.35	**0.76**	**0.66**
EucCos	**0.48**	0.73	0.64	**0.50**	0.71	**0.63**	0.24	**0.81**	**0.70**	0.44	0.78	0.68	**0.42**	**0.76**	**0.66**

to the concatenation (Conc) process used in the Stacked-RNNs architecture. As an illustrative case we show the TARN performance on the CAINS expression symptoms. Table 5.3 shows the results obtained by the TARN model over the different measures. The performance varies over the different measures. On average over the 4 CAINS-EXP symptoms, the EucCos measure leads to the best performance. Another measure that shows good performance is Mult. Although, EucCos and Mult are fixed measures with no learnable parameters, the following layers in the relation module are trainable and non-linear.

In the next experiment, we investigate the benefits of using segment-by-segment attention, and the entire videos (i.e. without clipping to make the videos have equal length) in our TARN architecture. To do so, we first compare the TARN model to another model that has no attention layer. The attention layer transforms the representations of one video to have the same number of representations as the other video. Subsequently, when the attention layer is not used, we use the video segmentation method applied in the Stacked-RNNs architecture, in order to pass a pair of videos with equal number of embeddings to the comparison layer. Second, we show how the TARN architecture with the attention layer performs when we use as input clipped videos with equal number of segments, instead of the entire videos. Third, we show how the TARN architecture performs when both the attention mechanism and the entire videos are used. In the three experiments, we use the EucCos measure in the comparison layer.

Table 5.4 shows the results obtained by the TARN architecture on treatment outcome estimation of the CAINS symptoms, using the different settings. By comparing the first and

Table 5.4: Performance of the TARN architecture at different settings.

Symptom	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			Average over all symptoms		
	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc
TARN (No attention, Clipped videos)	0.30	0.72	0.60	0.36	0.72	0.61	0.25	0.80	0.68	0.34	**0.79**	**0.68**	0.31	0.76	0.64
TARN (Attention, Clipped videos)	0.40	0.70	0.60	0.41	**0.76**	**0.66**	**0.26**	**0.81**	**0.70**	0.34	**0.79**	**0.68**	0.35	**0.77**	**0.66**
TARN (Attention, Entire videos)	**0.48**	**0.73**	**0.64**	**0.50**	0.71	0.63	0.24	**0.81**	**0.70**	**0.44**	0.78	**0.68**	**0.42**	0.76	**0.66**

second rows in Table 5.4, we can see that on average over the 4 CAINS-EXP symptoms, using attention improves the performance of the TARN architecture. Note that no difference in performance is observed for the quantity of speech symptom. A possible reason for that is that assessment interviews include asking the patients different questions throughout the interviews, and subsequently the amount of patients' speech is something that can be learned from the different video segments without alignment. Finally, by comparing the second and third rows, we can see that on average over the 4 CAINS-EXP symptoms, using the entire videos leads to a big improvement in the $F1_P$ score, at almost the same accuracy and $F1_N$ scores. Using the entire videos avoids losing any information from the videos, in addition it leads to more comparisons, which helps in improving the training process and reducing overfitting.

Symptom Severity Estimation (SSE). In order to test how SSE methods perform on treatment outcome estimation, these methods are applied for estimating the symptom severity before and after treatment independently, and then the results are compared so as to reach a conclusion on the treatment outcome. We report on three methods that have been used for SSE in schizophrenia, namely [135], [137], and SchiNet (proposed in Chapter 4). For a fair comparison, we have re-implemented [135], [137], and re-trained all methods using the 74 patients. We used the probabilities of the 11 detected AUs in the training and the LOSO protocol for training/testing. For each fold, we used 73 patients (146 videos) for training, and 1 patient (2 videos) for testing. For [137] and SchiNet, we tried different number of clusters or Gaussian components, and report the results of the best performing ones (12 clusters for [137] and 32 Gaussian components for SchiNet).

Table 5.5: Performance of the proposed architectures as well as other SSE methods on TOE for the **CAINS** expression symptoms.

Symptom	Facial Expression			Vocal Expression			Expressive Gestures			Quantity of Speech			Average over all symptoms		
	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc
Chance level	0.36	0.59	0.50	0.38	0.58	0.50	0.34	0.60	0.50	0.38	0.58	0.50	0.37	0.59	0.50
Tron *et al.* [135]	0.18	0.75	0.62	0.27	0.72	0.59	0.29	0.73	0.61	0.21	0.73	0.60	0.24	0.74	0.61
Tron *et al.* [137]	0.14	**0.80**	**0.67**	0.31	**0.77**	**0.66**	0.22	0.79	0.67	0.07	0.76	0.62	0.19	**0.78**	0.66
SchiNet	0.20	0.75	0.62	0.24	0.76	0.63	0.35	0.76	0.65	0.27	0.76	0.64	0.27	0.76	0.64
Stacked-RNNs	0.42	0.76	0.66	0.43	0.74	0.64	0.37	0.80	**0.70**	0.33	**0.79**	**0.68**	0.39	0.77	**0.67**
TARN-Mult	0.39	0.76	0.66	0.29	0.73	0.61	**0.42**	0.76	0.66	0.41	0.78	**0.68**	0.38	0.76	0.65
TARN-EucCos	**0.48**	0.73	0.64	**0.50**	0.71	0.63	0.24	**0.81**	**0.70**	**0.44**	0.78	**0.68**	**0.42**	0.76	0.66

Table 5.6: Performance of the proposed architectures as well as other SSE methods on TOE for the **PANSS** negative symptoms.

Symptom	Flat Affect			Poor Rapport			Lack of Spontaneity			Average over all symptoms		
	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc	$F1_P$	$F1_N$	Acc
Chance level	0.39	0.58	0.50	0.33	0.60	0.50	0.36	0.59	0.50	0.36	0.59	0.50
Tron *et al.* [135]	0.35	0.73	0.62	0.25	0.76	0.64	0.33	0.75	0.64	0.31	0.75	0.63
Tron *et al.* [137]	0.28	0.74	0.62	0.12	**0.83**	**0.71**	0.22	0.75	0.62	0.21	0.77	0.65
SchiNet	0.31	**0.76**	0.64	0.20	0.79	0.67	0.21	0.73	0.60	0.24	0.76	0.64
Stacked-RNNs	0.40	**0.76**	**0.66**	0.30	0.79	0.68	0.46	**0.80**	**0.71**	0.39	**0.78**	**0.68**
TARN-Mult	**0.50**	0.74	**0.66**	**0.36**	0.68	0.57	**0.48**	0.75	0.66	**0.45**	0.72	**0.68**
TARN-EucCos	0.45	0.73	0.64	0.28	0.78	0.66	0.36	**0.80**	0.69	0.36	0.77	0.66

Table 5.5 and 5.6 summarise the performance of the SSE methods on treatment outcome estimation for the CAINS and PANSS symptoms, respectively. Furthermore, we report the chance-level performance in Table 5.5 and 5.6. The SSE methods show relatively low performance when applied for estimating the treatment outcome. The reason for that is that often the change in negative symptoms during treatment is small [112], and falls within the error margin of the SSE methods.

Treatment Outcome Estimation (TOE). The performance of the two proposed architectures (Stacked-RNNs and TARN) on TOE for the CAINS and PANSS symptoms are shown in Table 5.5 and 5.6, respectively. For the TARN architecture, we report the performance when using both the EucCos and Mult measures in the comparison layer, as they are the best-performing ones in Table 5.3. Table 5.5 and 5.6 show that on average the Stacked-RNNs and TARN architectures outperform the SSE methods and the chance-level performance (better

$F1_P$ score at almost the same $F1_N$ and accuracy scores). Hence, building a network that analyses jointly a pair of patient's videos for TOE can extract more distinctive and related features to the treatment outcome than other networks that are trained specifically for SSE. Table 5.5 and 5.6 also show that the TARN architecture has better performance than Stacked-RNNs in 3 out of 4 CAINS-EXP symptoms (facial expression, vocal expression, and quantity of speech) and 1 PANSS-NEG symptom (flat affect). The two architectures have competitive performance in the remaining 3 symptoms (expressive gestures, poor rapport, and lack of spontaneity). The confusion matrices comparing the classification results of an SSE method (SchiNet) and the two proposed TOE methods (Stacked-RNNs and TARN) on TOE for the CAINS and PANSS symptoms, are shown in Figure 5.3 and 5.4, respectively. Note that in Figure 5.3 we use the TARN-EucCos model in the comparison, while in Figure 5.4 we use TARN-Mult, as the best performing measure in the TARN comparison layer varies from the CAINS to the PANSS scale.

In addition to the relatively good performance achieved by the TARN architecture in TOE, it has other pros compared to the Stacked-RNNs architecture. First, TARN uses an attention mechanism for aligning and comparing videos of variable temporal length. Second, TARN is easier to train and test as it is trained in an end-to-end fashion, while Stacked-RNNs is trained in two steps. Third, TARN does not have the exhaustive and highly computational feature selection method used in Stacked-RNNs. All of that inspired us to extend/modify the TARN architecture in [20] for addressing the problem of action recognition when a few training examples are available (i.e. few-shot learning), or when only a class description is given (i.e. zero-shot learning). TARN achieves the state-of-the-art results in few-shot action recognition, and very competitive performance in zero-shot action recognition.

5.4 Conclusion

In this chapter we propose two architectures for addressing directly the problem of treatment outcome estimation in schizophrenia. Both architectures exploit Deep Neural Networks for

Figure 5.3: The confusion matrices comparing the classification results of the SchiNet, Stacked-RNNs, and TARN methods on TOE for the CAINS-EXP symptoms.

learning differences in patients' behaviour, and in particular facial expressions. The first architecture (Stacked-RNNs) concatenates expressions/AUs extracted from a pair of videos recorded before and after treatment, and then pass the concatenated AUs to a deep network consisting of two stacked RNNs for learning local and global features over the pair of videos. The second architecture (TARN) includes an embedding module for encoding AUs probabilities over short segments of videos, and a relation module that uses an attention mechanism for

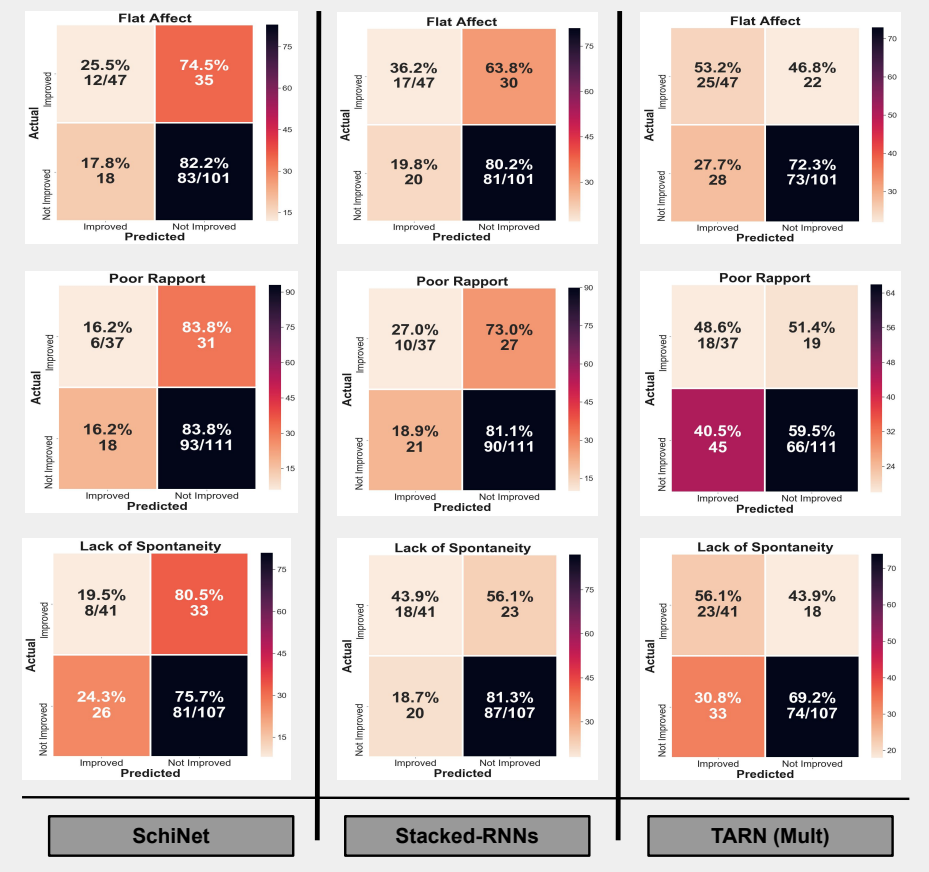


Figure 5.4: The confusion matrices comparing the classification results of the SchiNet, Stacked-RNNs, and TARN methods on TOE for the PANSS-NEG symptoms.

performing temporal alignment, and a deep network for learning a deep metric on the aligned representations at video segment level. The two architectures vary in two main aspects. First, Stacked-RNNs assumes that videos are aligned and equal in length, losing by that useful information, while TARN uses and aligns videos of different length. Second, Stacked-RNNs is trained in two steps, while TARN is trained end-to-end.

Symptom assessment interviews recorded in settings similar to real clinical ones are used in our analysis. Different negative symptoms from the PANSS and CAINS scales are estimated. The proposed architectures show better performance in Treatment Outcome Estimation (TOE), in comparison to other methods proposed for Symptom Severity Estimation (SSE). However, the SSE and TOE methods are complementary. More specifically, the SSE methods can be

used during patients' first sessions for diagnosis, while the TOE methods can be used during treatment/follow-up sessions for monitoring the change in symptoms levels. Experimental results also show that using attention and the entire videos in the analysis improve the TOE performance.

CHAPTER 6

Conclusion and discussion

This thesis focused on the problem of developing fully-automatic behaviour-based methodologies for diagnosis and treatment of schizophrenia, in settings that are similar to real clinical ones. This problem was divided into 3 subproblems. First, quantifying patients' non-verbal behaviour (facial expressions) during clinical interviews. Second, exploiting patients' expressions in diagnosing schizophrenia (i.e. estimating symptom severity). Third, comparing patients' expressions before and after treatment for determining the treatment outcome. Each of these subproblems was solved in one of the three main chapters in the thesis.

Literature review show that the methods that are typically used for analysing facial expressions of patients with schizophrenia work either on frontal views or in a specific environment – these methods are hard to perform well in real scenarios. In the first main chapter, two deep architectures were developed for analyzing facial expressions, these architectures were either trained using different datasets or in the wild, in order to be robust to the different recording conditions, found in real scenarios. The two architectures were trained to detect the activation of facial Action Units (AUs), the absence of which is expected to be informative in the diagnosis and treatment of schizophrenia.

Fusing different deep features (appearance, geometric, temporal) has not been explored yet in AUs detection, so in our first architecture we fused different deep models (CNNs, MLPs,

B-RNNs) together to exploit various kinds of information/features. The deep models were trained using 4 different datasets in order to increase the size of the training set, and include different recording conditions. Unlike other works in the literature that address the problem of AUs detection as a binary classification problem, where a different network is trained for each AU, and ignoring in this way informative correlations between AUs. In our first architecture, a multilabel classifier was employed in each of the deep models. We have two main contributions in the implemented multilabel classifier. First, a novel method was proposed for addressing the data imbalance problem in multilabel classification. That is, the cost term associated with each AU positive example was adjusted with the ratio of the negative to positive examples in the current batch. This method does not add any extra computational cost to the architecture. Second, the problem of threshold selection at the output neurons at test time was addressed by using two neurons for each class, one for positive activation while other for negative activation. During training, those output neurons were supervised with complementary information, and during testing, the maximum of the two neurons was chosen to represent the activation.

Experimental results show first that the different deep models (CNNs, MLPs, B-RNNs) perform significantly different in detecting AUs and the combined architecture is better than any single network. Second, adding each of the data-balancing and threshold-selection contributions improve the performance of our architecture. Finally, the proposed architecture achieved the state-of-the-art results on the BP4D dataset, and outperformed other works in the literature by a large margin.

The first architecture was trained using video data that were captured in controlled settings (available in the literature by that time). Later with the release of the EmotioNet dataset – a dataset that consists of facial images collected in the wild and annotated for different AUs – another architecture was developed and trained using the EmotioNet dataset as well as other datasets in the wild for the detection of AUs at more various recording settings (camera viewpoints and illumination levels). The reason a new architecture was developed is that

the datasets in the wild consist only of images, so no RNNs could be used in the second architecture. In addition, the facial images in these datasets have a wide range of head poses, so registering the face or extracting meaningful geometric features was quite challenging. Based on that, only CNNs were used in the second architecture, that is, deep pretrained CNNs (VGG-16) were refined for AUs detection. As the the number of annotated examples in the datasets vary immensely, a separate CNN was refined for each AU. Experiments showed promising detection results at different head poses and illumination levels.

At the end of this chapter, a qualitative and quantitative comparison was presented between the two AUs detection architectures, where strengths and weaknesses of each architecture were explained. The comparison is two-fold. First, the performance of the two architectures on AUs detection was compared in two settings, first on data captured in controlled settings and then in the wild. Second, the effect of each architecture on the performance of symptom severity estimation in schizophrenia was showed. For the first comparison, the first architecture showed better performance in detecting subtle AUs when tested in controlled settings, due to the various kinds of features used in it. On the other hand, the second architecture showed better performance in detecting AUs at different head poses and illumination levels (in the wild), as it was trained using images captured in the wild. For the second comparison, the second architecture showed better performance on estimating symptom severity. The reason for that is that the patients' interviews have a wide range of different recording conditions, and subsequently the second architecture leads to better analysis. At the end, we encourage the research community to address the limitations in the existing AU-annotated datasets, by collecting temporal datasets in the wild. In addition, we suggest for future work detecting the subjects' neutral face more accurately, and using it for normalizing the facial images used as input to the deep architecture – this can help the architecture learn more distinctive features away from the subjects' appearance differences – and subsequently improving the AUs detection performance.

In the second main chapter of the thesis, we addressed two limitations in the works that

used automatic behaviour analysis for studying and diagnosing schizophrenia. The **first** limitation is the use of structured interviews in the analysis. These interviews are different from the ones conducted in clinics and hospitals. Moving the analysis from controlled to real settings is quite challenging due to two main reasons. First, interviews recorded in realistic conditions have different recording conditions, and this can severely affect the performance of the facial expression analysis architectures. Second, interviews recorded in real scenarios can have different lengths, and classifiers like MLPs or SVMs work with features of fixed dimensionality. Hence, a fixed-length representation is required to be extracted from each of these varying-length videos. In this work, videos of professional-patient interviews, that were recorded in realistic conditions (i.e. varying illumination levels and camera viewpoints), were used in our analysis. In these interviews symptoms were assessed in a standardised way as they should/may be assessed in clinics and hospitals. Using these videos helped in moving from controlled contexts used in the literature to similar-to-real clinical settings. Furthermore, previous works used datasets consisting of a relatively small number of patients in diagnosing schizophrenia. In this work three times the highest number of patients used in other studies were analyzed.

The **second** limitation in the literature is the relatively low performance in estimating the severity of schizophrenia. One of the possible reasons behind that is the hand-crafted features used in the analysis – these features are difficult to generalize over different videos/patients, and subsequently can have implications on the performance of the regression models. The hand-crafted features have shown inferior performance in comparison to learned ones and in particular those learned by Deep Neural Networks. However, training deep networks with a large number of parameters (like CNNs) requires a large amount of data, and the number of patients available for the analysis in this kind of problems is relatively limited, due to the difficulty and the ethical issues in the collection of data depicting patients' behaviour. In addition, training deep temporal models like RNNs over long video sequences tend to suffer from the vanishing or exploding gradients problem. Subsequently, developing deep architectures that can learn distinctive features over limited number of patients and long sequences is quite

challenging.

In the second main chapter, a deep architecture, named SchiNet, was proposed for estimating symptom severity in schizophrenia. SchiNet uses quantified patients' facial expressions in symptom estimation. Given a video interview of a patient, the developed architectures in the first main chapter were first used for detecting patients' AUs (low-level features). Then, the AUs probabilities were used as input to a novel neural network (SchiNet) consisting of custom Gaussian Mixture Model and Fisher Vector layers for extracting a compact statistical feature vector over the whole video interview (high-level features), and a regression layer for symptom estimation. SchiNet has relatively limited number of trainable parameters, and can extract a fixed-length representation over long varying-length videos. Expression-related negative symptoms in two different assessment interviews, PANSS and CAINS, were estimated.

Our experimental results show many findings. First, significant correlations were found between the occurrence frequency of AUs and the severity of different symptoms in schizophrenia, resembling many of the correlations found across the literature. In addition, the found correlations confirm that symptom levels of patients with schizophrenia are expressed in the degree of their impairments in expression of emotion and social interaction. Second, our deep architecture (SchiNet) outperformed other state-of-the-art methods (that used hand-crafted features in their analysis) in 6 out of the 7 estimated symptoms. Third, several symptoms in the PANSS and CAINS interviews can be estimated with a mean absolute error less than one level. Finally, comparing the correlations between the automatic estimations and professional assessment (reaching at most to 0.46), to the correlations between assessments of different professionals that have annotated the NESS dataset [106] (equals to 0.85), shows that automatic estimation of symptom severity needs further improvement to reach human level performance. In order to improve the performance of symptom estimation, we suggest for future work extending the behaviour analysis to include body gestures and vocal expressions (besides facial expressions).

Many works in the literature have focused on classifying or estimating the severity of different mental illnesses, however, to the best of our knowledge, no works have addressed the problem of treatment outcome estimation in mental illnesses. In the last main chapter of the thesis, two deep architectures were proposed for addressing directly the problem of treatment outcome estimation in schizophrenia – more specifically, the proposed methods are aimed at determining whether specific symptoms have improved or not by analysing jointly two video interviews of the same patient, one before and one after the treatment. Both architectures first extract patient's facial expressions in the pair of videos, and then use it as input to a deep neural network. The first architecture (Stacked-RNNs) uses two stacked GRUs for learning local and global differences in patient's expressions/AUs over both videos. One GRU is used for learning behaviour differences over short video segments, while the other uses the segment-level features for learning global ones. The second architecture (TARN) includes two modules: the embedding module and the relation module. The embedding module includes a GRU for encoding the AUs probabilities in short segments of videos, while the relation module consists of an attention mechanism for aligning video segments, and a deep neural network for learning a deep metric for video comparison. The use of relation networks for estimating the treatment outcome is inspired by their success in data-limited problems (i.e. few-shot and zero-shot learning) [123]. The proposed Stacked-RNNs and TARN architectures have two main differences. First, Stacked-RNNs assumes videos are aligned and have equal length – this requires the clipping of long videos, and losing by that useful information. On the other hand, TARN assumes videos are not aligned and uses an attention mechanism for aligning videos of different lengths, avoiding by that any information loss. Second, Stacked-RNNs is trained in two steps, while TARN is trained in an end-to-end fashion.

In the analysis, symptom assessment interviews that were recorded in uncontrolled conditions and in settings that are similar to real clinical ones, were used. Different negative symptoms from the PANSS and CAINS interviews were estimated. Experimental results showed many findings. First, aligning patient's videos and using the entire videos (with no clipping) improved the performance of Treatment Outcome Estimation (TOE). Second, using the Euc-

Cos measure (which includes Euclidean distance and cosine similarity) in the TARN comparison layer leaded to better performance on TOE, in comparison to other measures like subtraction, concatenation, and neural network. Third, TARN achieved better TOE performance than Stacked-RNNs in many symptoms. Finally, Stacked-RNNs and TARN achieved better performance in TOE than other methods proposed for Symptom Severity Estimation (SSE) like SchiNet. The reason for that is that the change in negative symptoms is typically small [112], and falls within the error margin of the SSE methods. However, the SSE and TOE methods are complementary. Specifically, the SSE methods can be used during patients' baseline sessions for diagnosing symptoms, while the TOE methods can be used during treatment and follow-up sessions for estimating the change in symptoms levels.

Although the proposed methods for TOE show promising results, TOE still needs further improvement to reach the performance level required for medical applications. A key point for improving the performance is increasing the training set size by including more patients. In our future work, we plan to extend behaviour analysis to include body gestures and vocal expressions, in addition to improving the AUs detection by moving from the static to the temporal analysis in the wild. Furthermore, we will use the quantified behaviour in estimating not only if symptoms have improved or not, but also the exact change in symptom levels.

APPENDIX **A**

Symptom assessment interviews in schizophrenia

Contents

A.1 Positive and negative syndrome scale

Positive and Negative Syndrome Scale (PANSS) consists of a total of 30 symptoms divided into 3 scales: negative, positive, and general psychopathology [75]. Out of the 30 symptoms, 7 are grouped to form the negative scale, 7 form the positive scale, and the remaining 16 symptoms form the general psychopathology scale. Each symptom in the PANSS is rated between 1 (absent) and 7 (extreme), according to the criteria provided in [75]. The total score of each scale is the summation of the ratings of the scale symptoms. Hence, the total score of the positive and negative scale ranges between 7-49, and between 16-112 for the general psychopathology scale.

Negative symptoms indicate a lack of normal mental functions like motivation, concentration, or/and expression. The lack in expression is marked by a reduction in patients' behaviour.

The negative symptoms include:

1. Blunted/Flat affect.
2. Emotional withdrawal.
3. Poor rapport.
4. Passive/Apathetic social withdrawal.
5. Difficulty in abstract thinking.
6. Lack of spontaneity and flow of conversation.
7. Stereotyped thinking.

Positive symptoms refer to thoughts or behaviour that are usually not seen in healthy people. The positive symptoms include:

1. Delusions.
2. Conceptual disorganization.
3. Hallucinations.
4. Excitement.
5. Grandiosity.
6. Suspiciousness/Persecution.
7. Hostility.

The general psychopathology scale includes the following symptoms:

1. Somatic concern.
2. Anxiety.
3. Guilt feelings.
4. Tension.
5. Mannerisms and posturing.
6. Depression.

7. Motor retardation.
8. Uncooperativeness.
9. Unusual thought content.
10. Disorientation.
11. Poor attention.
12. Lack of judgment and insight.
13. Disturbance of volition.
14. Poor impulse control.
15. Preoccupation.
16. Active social avoidance.

A.2 Clinical assessment interview for negative symptoms

Clinical Assessment Interview for Negative Symptoms (CAINS) measures severity of negative symptoms in patients with schizophrenia [61]. Unlike the PANSS interview that reports a single score for negative symptoms, CAINS consists of 2 negative scales that are rated separately: motivation and pleasure and expression. The motivation and pleasure scale has 9 symptoms, and the expression scale has 4 symptoms. Each symptom in CAINS has a value between 0 and 4 (0=no impairment and 4=severe impairment).

The motivation and pleasure scale measures impairment in motivation for social relationships, school/work activities and recreation, and includes the following symptoms:

1. Motivation for close family/spouse/partner relationships.
2. Motivation for close friendships/romantic relationships.
3. Frequency of pleasurable social activities – past week.
4. Frequency of expected pleasure from social activities – next week.
5. Motivation for work and school activities.
6. Frequency of expected pleasure from work and school activities – next week.
7. Motivation for recreational activities.

8. Frequency of pleasurable recreational activities – past week.
9. Frequency of expected pleasure from recreational activities – next week.

The expression scale measures impairment in expression of emotion and speech. The rating of the expression symptoms depends on observed emotional behaviour throughout the whole interview. The expression symptoms include:

1. Facial expression.
2. Vocal expression.
3. Expressive gestures.
4. Quantity of speech.

Bibliography

[1] L. A. Abel, L. Friedman, J. Jesberger, A. Malki, and H. Meltzer. Quantitative assessment of smooth pursuit gain and catch-up saccades in schizophrenia and affective disorders. *Biological psychiatry*, 29(11):1063–1072, 1991. 18

[2] S. Alghowinem, R. Goecke, J. F. Cohn, M. Wagner, G. Parker, and M. Breakspear. Cross-cultural detection of depression from nonverbal behaviour. 19

[3] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, G. Parker, et al. From joyous to clinically depressed: Mood detection using spontaneous speech. In *FLAIRS Conference*, 2012. 19

[4] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Hyett, G. Parker, and M. Breakspear. Multimodal depression detection: fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, 2016. 19, 28

[5] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear. Eye movement analysis for depression detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 4220–4224. IEEE, 2013. 19, 28

[6] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 283–288. IEEE, 2013. 19, 28

[7] C. Alvino, C. Kohler, F. Barrett, R. E. Gur, R. C. Gur, and R. Verma. Computerized measurement of facial expression of emotions in schizophrenia. *Journal of neuroscience methods*, 163(2):350–361, 2007. 3, 15, 16, 17

[8] N. C. Andreasen. Scale for the assessment of negative symptoms (sans). *The British Journal of Psychiatry*, 1989. 17

[9] A. Andrew, M. Knapp, P. R. McCrone, M. Parsonage, and M. Trachtenberg. Effective interventions in schizophrenia: the economic case. 2012. 2

[10] S. Annen, P. Roser, and M. Brüne. Nonverbal behavior during clinical interviews: similarities and dissimilarities among schizophrenia, mania, and depression. *The Journal of nervous and mental disease*, 200(1):26–32, 2012. 72, 73

[11] G. Antipov, S.-A. Berrani, N. Ruchaud, and J.-L. Dugelay. Learned vs. hand-crafted features for pedestrian gender recognition. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1263–1266. ACM, 2015. 23, 58

[12] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 89

[13] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic 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. 23, 24, 25, 44, 45

[14] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 5, pages 53–53. IEEE, 2003. 3

[15] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez. Emotionet challenge: Recognition of facial expressions of emotion in the wild. *arXiv preprint arXiv:1703.01210*, 2017. 50

[16] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012. 38

[17] M. Bishay, P. Palasek, S. Priebe, and I. Patras. Schinet: Automatic estimation of symptoms of schizophrenia from facial behaviour analysis. *arXiv preprint arXiv:1808.02531*, 2018. 29, 57

[18] M. Bishay and I. Patras. Fusing multilabel deep networks for facial action unit detection. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 681–688. IEEE, 2017. 29, 47

[19] M. Bishay, S. Priebe, and I. Patras. Can automatic facial expression analysis be used for treatment outcome estimation in schizophrenia? In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1632–1636. IEEE, 2019. 80

[20] M. Bishay, G. Zoumpourlis, and I. Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019. 80, 82, 97

[21] M. Brüne, C. Sonntag, M. Abdel-Hamid, C. Lehmkämper, G. Juckel, and A. Troisi. Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders. *The Journal of nervous and mental disease*, 196(4):282–288, 2008. 2, 72, 73

[22] H. Cai, J. Han, Y. Chen, X. Sha, Z. Wang, B. Hu, J. Yang, L. Feng, Z. Ding, Y. Chen, et al. A pervasive approach to eeg-based depression detection. *Complexity*, 2018, 2018. 15

[23] S. Canavan, M. Chen, S. Chen, R. Valdez, M. Yaeger, H. Lin, and L. Yin. Combining gaze and demographic feature descriptors for autism classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3750–3754. IEEE, 2017. 22, 23

[24] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013. 16

[25] N. V. Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 853–867. Springer, 2005. 32, 49

[26] K. Cho, B. Van Merriënboer, C. Gulcehre, et al. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 86

[27] Y. Cho, N. Bianchi-Berthouze, and S. J. Julier. Deepbreath: Deep learning of breathing patterns for automatic stress recognition using low-cost thermal imaging in unconstrained settings. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 456–463. IEEE, 2017. 15

[28] Y. Cho, S. J. Julier, and N. Bianchi-Berthouze. Instant stress: Detection of perceived mental stress through smartphone photoplethysmography and thermal imaging. *JMIR mental health*, 6(4):e10140, 2019. 15

[29] W.-S. Chu, F. De la Torre, and J. F. Cohn. Modeling spatial and temporal cues for multi-label facial action unit detection. *arXiv preprint arXiv:1608.00911*, 2016. 40, 45, 46

[30] W.-S. Chu, F. De la Torre, and J. F. Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017. 23, 24

[31] J. F. Cohn and F. De la Torre. Automated face analysis for affective. *The Oxford handbook of affective computing*, page 131, 2014. 64

[32] J. F. Cohn and P. Ekman. Measuring facial action. *The new handbook of methods in nonverbal behavior research*, pages 9–64, 2005. 3

[33] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. La Torre. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–7. IEEE, 2009. 15, 18

[34] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001. 18

[35] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015. 15

[36] C. Darwin. *The expression of the emotions in man and animals*. John Murray, 1872. 3

[37] P. Davison, C. Frith, P. Harrison-Read, and E. Johnstone. Facial and other non-verbal communicative behaviour in chronic schizophrenia. *Psychological medicine*, 26(04):707–713, 1996. 2

[38] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. 65

[39] S. Dieleman, J. Schlüter, C. Raffel, E. Olson, S. K. Sønderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, et al. Lasagne: first release. *Zenodo: Geneva, Switzerland*, 3, 2015. 74, 93

[40] S. Dimic, C. Wildgrube, R. McCabe, I. Hassan, T. R. Barnes, and S. Priebe. Non-verbal behaviour of patients with schizophrenia in medical consultations–a comparison with depressed patients and association with symptom levels. *Psychopathology*, 43(4):216–222, 2010. 2, 29, 72, 73

[41] Z. Du, W. Li, D. Huang, and Y. Wang. Bipolar disorder recognition via multi-scale discriminative audio temporal representation. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 23–30. ACM, 2018. 22

[42] G.-B. Duchenne. Mécanisme de la physionomie humaine: où, analyse électrophysiologique de l'expression des passions. *Archives générales de médecine*, pages 29–47, 1862. 3

[43] Z. Dvey-Aharon, N. Fogelson, A. Peled, and N. Intrator. Schizophrenia detection and classification by advanced analysis of eeg recordings using a single electrode approach. *PloS one*, 10(4):e0123033, 2015. 15

[44] P. Ekman and E. L. Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997. 4, 63

[45] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5562–5570, 2016. viii, ix, x, 46, 48, 49, 50, 55

[46] L. A. Fairbanks, M. T. McGuire, and C. J. Harris. Nonverbal interaction of patients and therapists during psychiatric interviews. *Journal of abnormal psychology*, 91(2):109, 1982. 14

[47] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern recognition*, 36(1):259–275, 2003. 15

[48] L. Fossi, C. Faravelli, and M. Paoli. The ethological approach to the assessment of depressive disorders. *The Journal of nervous and mental disease*, 172(6):332–341, 1984. 18

[49] H. . C. P. D. General. Improving the mental health of the population: Towards a strategy on mental health for the european union. *Technical report, European Union*, 2005. 1

[50] 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. IEEE, 2015. 23, 24, 25, 47

[51] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014. 14

[52] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and Vision Computing*, 32(10):641–647, 2014. 21

[53] Y. Gong, H. Yatawatte, C. Poellabauer, S. Schneider, and S. Latham. Automatic autism spectrum disorder detection using everyday vocalizations captured by smart devices. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 465–473. ACM, 2018. 15

[54] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013. 37

[55] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis. Deep learning based facs action unit occurrence and intensity estimation. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 6, pages 1–5. IEEE, 2015. 23, 24, 25, 44, 45, 47

[56] W. W. Hale III, J. H. Jansen, A. L. Bouhuys, J. A. Jenner, and R. H. van den Hoofdakker. Non-verbal behavioral interactions of depressed patients with partners and strangers: The role of behavioral social support and involvement in depression persistence. *Journal of affective disorders*, 44(2):111–122, 1997. 18

[57] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2):237–256, 2011. 16, 17, 72

[58] J. Hamm, A. Pinkham, R. C. Gur, R. Verma, and C. G. Kohler. Dimensional information-theoretic measurement of facial emotion expressions in schizophrenia. *Schizophrenia research and treatment*, 2014, 2014. 15, 16, 17

[59] B.-C. Ho, P. Nopoulos, M. Flaum, S. Arndt, and N. C. Andreasen. Two-year outcome in first-episode schizophrenia: predictive value of symptoms for quality of life. *Focus*, 155(1):1196–137, 2004. 2

[60] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 86

[61] W. P. Horan, A. M. Kring, R. E. Gur, S. P. Reise, and J. J. Blanchard. Development and psychometric validation of the clinical assessment interview for negative symptoms (cains). *Schizophrenia research*, 132(2):140–145, 2011. 58, 59, 68, 82, 91, 92, 110

[62] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. 2008. 49

[63] D. Iter, J. Yoon, and D. Jurafsky. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146, 2018. 15

[64] S. Jaiswal and M. Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 23, 24, 25, 35, 44, 45

[65] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley. Automatic detection of adhd and asd from expressive behaviour in rgbd data. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 762–769. IEEE, 2017. 15, 22, 23

[66] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh. Automatic depression scale prediction using facial expression dynamics and regression. In *Proceedings of the*

4th International Workshop on Audio/Visual Emotion Challenge, pages 73–80. ACM, 2014. 20, 28

[67] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):668–680, 2017. 20, 21

[68] Y. Jang, H. Gunes, and I. Patras. Smilenet: Registration-free smiling face detection in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1581–1589, 2017. 46, 48, 85

[69] J. Joshi, A. Dhall, R. Goecke, M. Breakspear, and G. Parker. Neural-net classification for spatio-temporal descriptor based depression analysis. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2634–2638. IEEE, 2012. 18, 19

[70] J. Joshi, A. Dhall, R. Goecke, and J. F. Cohn. Relative body parts movement for automatic depression analysis. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 492–497. IEEE, 2013. 18, 19

[71] J. Joshi, R. Goecke, G. Parker, and M. Breakspear. Can body expressions contribute to automatic depression analysis? In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–7. IEEE, 2013. 18, 19

[72] H. Jung, S. Lee, S. Park, I. Lee, C. Ahn, and J. Kim. Deep temporal appearance-geometry network for facial expression recognition. *arXiv preprint arXiv:1503.01532*, 2015. 23, 36, 37

[73] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. *depression*, 1:1, 2014. 19

[74] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al. Emonets: Multimodal

deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, pages 1–13, 2015. 38

[75] S. R. Kay, A. Flszbein, and L. A. Opfer. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261, 1987. 17, 58, 59, 68, 82, 91, 92, 108

[76] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 93

[77] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015. 90

[78] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010. 23

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 23, 35, 47, 58

[80] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 18

[81] M. Lavelle, P. G. Healey, and R. McCabe. Is nonverbal communication disrupted in interactions involving patients with schizophrenia? *Schizophrenia bulletin*, 39(5):1150–1158, 2012. 2, 29

[82] M. Lavelle, P. G. Healey, and R. McCabe. Nonverbal behavior during face-to-face social interaction in schizophrenia: a review. *The Journal of nervous and mental disease*, 202(1):47–54, 2014. 2

[83] Q. V. Le, N. Jaitly, and G. E. Hinton. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*, 2015. 37

[84] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3361–3368. IEEE, 2011. 23, 58

[85] A. F. Leentjens, S. M. Wielaert, F. van Harskamp, and F. W. Wilmink. Disturbances of affective prosody in patients with schizophrenia; a cross sectional study. *Journal of Neurology, Neurosurgery & Psychiatry*, 64(3):375–378, 1998. 2

[86] W. Li, F. Abtahi, Z. Zhu, and L. Yin. Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection. *arXiv preprint arXiv:1702.02925*, 2017. 40, 45, 46

[87] R. B. Lipton, S. Levin, and P. S. Holzman. Horizontal and vertical pursuit eye movements, the oculocephalic reflex, and the functional psychoses. *Psychiatry Research*, 3(2):193–203, 1980. 18

[88] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 61, 63, 83

[89] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. viii, ix, x, 48, 49, 50, 55

[90] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011. ix, 39, 40

[91] J. Mackintosh, R. Kumar, and T. Kitamura. Blink rate in psychiatric illness. *The British Journal of Psychiatry*, 143(1):55–57, 1983. 14

[92] J. Mackintosh, R. Kumar, and T. Kitamura. Blink rate in psychiatric illness. *The British Journal of Psychiatry*, 143(1):55–57, 1983. 18

[93] M. K. Mandal, R. Pandey, and A. B. Prasad. Facial expressions of emotions and schizophrenia: A review. *Schizophrenia bulletin*, 24(3):399, 1998. 2

[94] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic. Automatic analysis of facial actions: A survey. *IEEE Transactions on Affective Computing*, 2017. 15, 64

[95] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. ix, 39, 40

[96] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2011. ix, 39, 40

[97] H. Meng, D. Huang, H. Wang, H. Yang, M. Ai-Shuraifi, and Y. Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30. ACM, 2013. 20

[98] H.-J. Möller. Clinical evaluation of negative symptoms in schizophrenia. *European Psychiatry*, 22(6):380–386, 2007. 2

[99] D. Murphy and J. Cutting. Prosodic comprehension and expression in schizophrenia. *Journal of Neurology, Neurosurgery & Psychiatry*, 53(9):727–730, 1990. 2

[100] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814, 2010. 36

[101] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 31

[102] P. Palasek and I. Patras. Discriminative convolutional fisher vector network for action recognition. *arXiv preprint arXiv:1707.06119*, 2017. 63, 65, 66

[103] A. Pampouchidou, P. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis. Automatic assessment of depression based on visual cues: A systematic review. *IEEE Transactions on Affective Computing*, 2017. 15

[104] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Nov. 2016. 89

[105] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza. Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 49–55, 2014. 28

[106] S. Priebe, M. Savill, T. Wykes, R. Bentall, U. Reininghaus, C. Lauber, S. Bremner, S. Eldridge, and F. Röhricht. Effectiveness of group body psychotherapy for negative symptoms of schizophrenia: multicentre randomised controlled trial. *The British Journal of Psychiatry*, pages bjp–bp, 2016. 59, 60, 79, 90, 91, 105

[107] S. S. Rajagopalan, O. R. Murthy, R. Goecke, and A. Rozga. Play with me—measuring a child's engagement in a social interaction. In *2015 11th IEEE International Conference*

and Workshops on Automatic Face and Gesture Recognition (FG), volume 1, pages 1–8. IEEE, 2015. 23

[108] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, et al. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13. ACM, 2018. 22

[109] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kocisky, and P. Blunsom. Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*, 2016. 89

[110] S. Saha, D. Chant, J. Welham, and J. McGrath. A systematic review of the prevalence of schizophrenia. *PLoS medicine*, 2(5):e141, 2005. 1

[111] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013. 64, 65, 66, 74

[112] M. Savill, C. Banks, H. Khanom, and S. Priebe. Do negative symptoms of schizophrenia change over time? a meta-analysis of longitudinal data. *Psychological medicine*, 45(8):1613–1627, 2015. 2, 8, 12, 58, 59, 74, 81, 96, 107

[113] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency. Automatic behavior descriptors for psychological disorder analysis. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2013. 21

[114] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997. 37

[115] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 57–63. ACM, 2014. 20

[116] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8, 47

[117] J. Snell, K. Swersky, and R. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 2017. 90

[118] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011. 37

[119] F. Song, X. Tan, X. Liu, and S. Chen. Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recognition*, 47(9):2825–2838, 2014. viii, ix, x, 48, 49, 50, 55

[120] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 36, 93

[121] S. R. Staugaard. Threatening faces and social anxiety: a literature review. *Clinical psychology review*, 30(6):669–690, 2010. 14

[122] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. Automatic nonverbal behavior indicators of depression and ptsd: the effect of gender. *Journal on Multimodal User Interfaces*, 9(1):17–29, 2014. 21

[123] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 28, 106

[124] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 90

[125] M. Suwa. A preliminary note on pattern recognition of human emotional expression. In *Proc. of The 4th International Joint Conference on Pattern Recognition*, pages 408–410, 1978. 4

[126] Z. S. Syed, K. Sidorov, and D. Marshall. Automated screening for bipolar disorder from audio/visual modalities. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 39–45. ACM, 2018. 22

[127] T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016. 74, 93

[128] Y.-L. Tian, T. Kanade, and J. F. Cohn. Facial expression analysis. In *Handbook of face recognition*, pages 247–275. Springer, 2005. 15

[129] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 50, 93

[130] A. Torralba, A. A. Efros, et al. Unbiased look at dataset bias. In *CVPR*, volume 1, page 7. Citeseer, 2011. 41

[131] F. Trémeau. A review of emotion deficits in schizophrenia. *Dialogues in clinical neuroscience*, 8(1):59, 2006. 2

[132] A. Troisi. Ethological research in clinical psychiatry: the study of nonverbal behavior during interviews. *Neuroscience & Biobehavioral Reviews*, 23(7):905–913, 1999. 4, 29, 72

[133] A. Troisi, E. Pompili, L. Binello, and A. Sterpone. Facial expressivity during the clinical interview as a predictor functional disability in schizophrenia. a pilot study. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 31(2):475–481, 2007. 72, 73

[134] A. Troisi, G. Spalletta, and A. Pasini. Non-verbal behaviour deficits in schizophrenia: an ethological study of drug-free patients. *Acta Psychiatrica Scandinavica*, 97(2):109–115, 1998. 2, 14

[135] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 72–81. Springer, 2015. 3, 16, 17, 28, 57, 72, 75, 76, 77, 80, 95, 96

[136] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Differentiating facial incongruity and flatness in schizophrenia, using structured light camera data. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*, pages 2427–2430. IEEE, 2016. 16, 17

[137] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall. Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 220–223. IEEE, 2016. 3, 16, 17, 57, 75, 76, 77, 80, 95, 96

[138] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2014. 19

[139] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 3–10. ACM, 2013. 19

[140] M. F. Valstar, T. Almaev, J. M. Girard, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. 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, 2015. 23, 39, 44, 45

[141] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170. ACM, 2006. 64

[142] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of neuroscience methods*, 168(1):224–238, 2008. 3, 16, 17

[143] P. Wang, C. Kohler, F. Barrett, R. Gur, and R. Verma. Quantifying facial expression abnormality in schizophrenia by combining 2d and 3d features. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 16, 17

[144] S. Wang and J. Jiang. A compare-aggregate model for matching text sequences. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 89, 93

[145] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014. 20, 28

[146] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta. Vocal and facial biomarkers of depression based on motor incoordination and timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. ACM, 2014. 64

[147] E. Worswick, S. Dimic, C. Wildgrube, and S. Priebe. Negative symptoms and avoidance of social interaction: A study of non-verbal behaviour. *Psychopathology*, 2017. 2, 14, 29, 72, 73

[148] X. Xing, B. Cai, Y. Zhao, S. Li, Z. He, and W. Fan. Multi-modality hierarchical recall based on gbdts for bipolar disorder classification. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 31–37. ACM, 2018. 22

[149] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013. 35

[150] L. Yang, Y. Li, H. Chen, D. Jiang, M. C. Oveneke, and H. Sahli. Bipolar disorder recognition with histogram features of arousal and body gestures. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 15–21. ACM, 2018. 22

[151] Y. Yang, C. Fairbairn, and J. F. Cohn. Detecting depression severity from vocal prosody. *Affective Computing, IEEE Transactions on*, 4(2):142–150, 2013. 19

[152] A. Yüce, H. Gao, and J.-P. Thiran. Discriminant multi-label manifold embedding 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. 23, 24, 44, 45

[153] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014. ix, 39, 40

[154] C. C. L. Zhanpeng Zhang, Ping Luo and X. Tang. From facial expression recognition to interpersonal relation prediction. In *arXiv:1609.06426v2*, 2016. ix, x, 48, 49, 50

[155] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007. 18

[156] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang. Joint patch and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 2207–2216, 2015. 23, 31, 40, 45, 46

[157] K. Zhao, W.-S. Chu, and H. 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. 23, 24, 25, 35, 40, 45, 46

[158] X. Zhu and D. 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, 2012. 34