

# SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis

Mina Bishay, Petar Palasek, Stefan Priebe, and Ioannis Patras

**Abstract**—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. In this paper, we use videos of professional-patient interviews, in which symptoms were assessed in a standardised way as they should/may be assessed in practice, and which were recorded in realistic conditions (i.e. varying illumination levels and camera viewpoints) at the patients' homes or at mental health services. We automatically analyse the facial behaviour of 91 out-patients – this is almost 3 times the number of patients in other studies – and propose SchiNet, a novel neural network architecture that estimates expression-related symptoms in two different assessment interviews. We evaluate the proposed SchiNet for patient-independent prediction of symptoms of schizophrenia. Experimental results show that some automatically detected facial expressions are significantly correlated to symptoms of schizophrenia, and that the proposed network for estimating symptom severity delivers promising results.

**Index Terms**—Automatic analysis, Non-verbal behaviour, Facial expression, Health care, Schizophrenia, Negative symptoms, Gaussian mixture model, Fisher vector.

## 1 INTRODUCTION

SCHIZOPHRENIA is a severe mental illness affecting not only the patients, but also their families and the society as a whole. Patients with schizophrenia often show impairment in the expression of emotion and speech in comparison to non-patients [1] – this is manifested in their facial expression [2], vocal expression [3], [4], and expressive gestures [5], [6]. Patients can also show impairment in the non-verbal behaviour that invites social interaction during clinical and nonclinical interviews [7]. Non-verbal behaviour was found to change during interviews according to symptom severity [8], [9], [10]. For instance, patients with high symptom severity tend to avoid interaction by nodding less, smiling less, and looking less at the interviewer [8]. Such impairments present valuable information for the psychiatrists, as they can be used for diagnosis and symptom assessment. However, behaviour analysis is time-consuming in research settings and subjective in clinical settings – this calls for the development of automatic analysis tools.

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) [11], [12], [13], [14]. 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, such as listening to life vignettes, or answering emotional questions. In addition, the tools that are typically used/proposed for AFEA perform well primarily in a specific, controlled environment that is hard to be replicated in clinics and hospitals. Furthermore, the methods proposed up to now for diagnosing schizophrenia rely on conventional hand-crafted features [12], [14]. Across different contexts, hand-crafted features have shown inferior performance in comparison to learned ones and in particular those learned by Deep Neural Networks [15], [16], [17].

In this paper, we move from controlled environments to similar-to-real-life settings and use professional-patient interviews of symptom assessment. More specifically, we use research 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 [18]. The interviews were recorded either at the patients' homes or at the premises of mental health services across the UK. The collected videos have a wide range of camera viewpoints and illumination levels that are representative of the variety of settings found in clinics. We used interviews of 91 out-patients – this is almost 3 times the highest number of patients used in other studies.

In order to automatically analyse the videos, we propose a Deep Neural Network (DNN) architecture, called SchiNet, that analyses facial expressions and estimates symptoms of schizophrenia that are related to them. The proposed SchiNet is patient-independent and consists of two main stages. In the first stage, different DNNs are used for detecting patients' facial expressions, such as smiles and activations of facial muscles at each frame (low-level features). At the second stage, a DNN consisting of a Gaussian Mixture Model (GMM) and Fisher Vector (FV) layers

- *Mina Bishay and Ioannis Patras are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK. E-mail: {m.a.t.bishay,i.patras}@qmul.ac.uk.*
- *Petar Palasek was with the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK. He is currently working as a research scientist at MindVisionLabs LTD. E-mail: p.palasek@qmul.ac.uk.*
- *Stefan Priebe is with the Unit for Social and Community Psychiatry (WHO Collaborating Centre for Mental Health Service Development), Newham Centre for Mental Health, Queen Mary University of London, London, UK. E-mail: s.priebe@qmul.ac.uk.*

for extracting a compact statistical feature vector over the whole video interview (high-level features), and b) a regression layer is used for symptom estimation. The different sub-networks are first trained in stages and then are refined in an end-to-end fashion.

The proposed network has been trained in a person-independent manner to predict expression-related symptoms from two commonly-used assessment interviews; Positive and Negative Syndrome Scale (PANSS) [19], and Clinical Assessment Interview for Negative Symptoms (CAINS) [20]. Experimental results show that training the Facial Expression Analysis sub-network (stage 1) “in the wild” delivers better performance on symptom severity estimation in comparison to another state-of-the-art method [21] that is trained using data captured in a controlled environment. 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.

The main contributions of our work are two-fold:

- 1) We move from controlled contexts to settings that are similar to real life ones, where we analyse symptom assessment interviews of almost three times the number of patients used in previous studies.
- 2) We propose a fully-automatic deep learning approach for estimating expression-related symptoms of schizophrenia in two different assessment interviews, namely PANSS and CAINS.

The rest of the paper is organized as follows: In Section 2, we review the related literature in analysing and diagnosing schizophrenia from facial expressions. In Section 3, we introduce the clinical dataset that we used in the analysis. In Section 4, we present the proposed SchiNet for estimating symptoms of schizophrenia. Finally, in Section 5 and Section 6 we give the experimental results and the conclusions, respectively.

## 2 RELATED WORK

Some psychiatric researches are concerned with the relation between schizophrenia and the patients’ non-verbal behaviour [6], [8], [9], [10], [22]. 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. For this reason, in the last few years there has been a growing interest in the application of Automatic Facial Expression Analysis (AFEA) methods for studying and diagnosing schizophrenia. In this section, we review related works 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 work we refer to the former as Penn-dataset, and the later as HUJI-dataset. The Penn-dataset consists of videos and images that are collected at two different sessions. In the first session, patients with schizophrenia and healthy controls are asked to express basic emotions at 3 different intensities. In the second session, they are 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 [11], [13], [23], [24], [25], 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 [12], [14], [26].

**AFEA methods.** Different methods have been used/proposed in the literature for analysing patients’ facial behaviour. In [11], Alvino *et al.* detected static emotional expressions by measuring a deformation between a neutral face and a face with expression, which was then classified using an SVM classifier. In [13], 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 [24], [25], Hamm *et al.* moved from analysing basic emotions to detection of 15 Action Units (AUs) at every frame of the sequence. The 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. As we will show, such methods are not suitable for analysis of non-verbal behaviour in uncontrolled conditions, such as professional-patient interviews, as they are not robust to variabilities in recording factors such as camera viewpoint and illumination levels. Similar results are reported in other studies: For example, [27] reports that the commercial 3D facial analysis tool used for detecting 23 AUs in [12], [14], [26], 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 extracted from facial expression analysis investigating the existence of differences between them. In addition, correlations between these features and flatness and inappropriateness symptoms in the SANS scale [28] were tested. Various features were extracted in these studies. In [11], [13], the average probability of 4 emotions and neutral expression were calculated. In [23], 2D geometric features and 3D curvature features were used in the comparison. In [24], features as frequency of some single and combined AUs were extracted, while in [25] 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 [26] 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.* [12] [14] go beyond studying the differences in behaviour between patients and healthy controls, and more specifically, use automatic analysis of facial behaviour for diagnosis and severity estimation of some PANSS symptoms

(especially flat affect). In these studies, different features were extracted and used with a two-step SVM based algorithm for the diagnosis and symptom estimation. In [12], features related to the intensity and dynamics of each AU (e.g. frequency, activation length, change ratio) were extracted, while in [14], clustering analysis was used over all AUs for extracting 3 flatness-related features, richness (number of facial-clusters appeared), typicality (the similarity to prototype), and cluster distribution (the activation frequency of different clusters).

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 problems of the diagnosis and symptom estimation in schizophrenia. Second, the datasets used in these works contain a relatively small number of patients and were recorded while the patients were performing controlled tasks. Third, the tools used/proposed for facial expression analysis work either on frontal views or in a specific environment. Finally, all the features used in the diagnosis and symptom estimation in schizophrenia are hand-crafted ones – this can have implications on the performance of the regression/classification model. By contrast, we use video recordings of 91 patients in conditions that are similar to realistic symptom assessment interviews. In addition, we use statistical deep features for estimating expression-related symptoms in two different assessment interviews, PANSS and CAINS. All of the networks that we use, including the first stage that analyses facial expressions, are trained with data “in the wild”, in order for them to be robust to different recording conditions.

### 3 CLINICAL DATASET OF SCHIZOPHRENIA

In this work we use a dataset called “NESS”, that was collected for studying the effectiveness of group body psychotherapy on negative symptoms of schizophrenia [29]. The participants in this study 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 [18].

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 [19] including negative, positive and general psychopathology symptoms, and CAINS [20] 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 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  $1920 \times 1080$ . Out of the 275 patients, 110 accepted to be recorded at BaseLine, 93 at End of Treatment, and 69 at 6 Months Follow-Up. Since the

focus of this paper 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. More information about the dataset can be found in [29].

## 4 PROPOSED ARCHITECTURE

### 4.1 Overview

In this section we present a deep architecture, named SchiNet, for estimating the severity of symptoms of schizophrenia from videos depicting the non-verbal behaviour of patients. Figure 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 [30] and a robust face detector [31]. At the second stage, the face regions are cropped and passed to a bank of Deep Neural Networks (DNNs), each of which detects a certain facial expression or the activation of a certain facial Action Unit. 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 4 stages, as shown in Figure 1(b). At each stage, a different cost is optimised. In the first stage, the DNNs that detect the activation of facial expressions at each frame are trained in a supervised manner on datasets annotated with the corresponding labels. At the second stage, the network that extracts video-based representations is trained in an unsupervised manner, taking as input the sequence of the outputs of the first network when applied to the professional-patient interviews. More specifically, the distribution of the expression 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 third stage the FV representations are used 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 [32], we refine the GMM, the FV and the first regression layer in an end-to-end fashion using a discriminative cost. Finally, in the fourth 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 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. We

a) SchiNet architecture

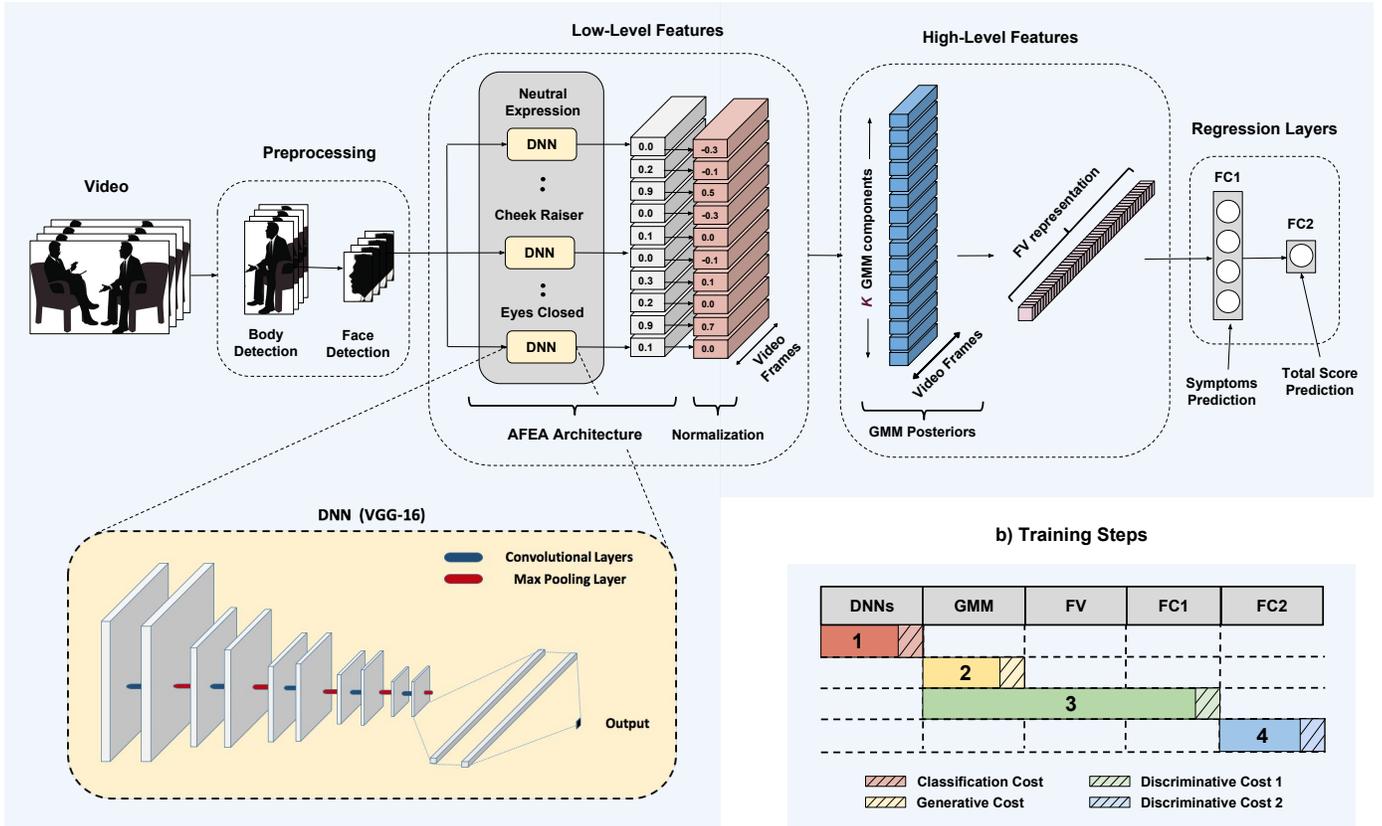


Fig. 1. (a) The proposed SchiNet for symptom severity estimation in schizophrenia. The input is a recorded video interview of a patient during his/her symptom assessment, and the outputs are the estimated values for the expression-related symptoms and the total scale/symptoms score. Feature extraction is done over two stages, first, the video is encoded by patient facial expressions, then a compact statistical feature vector is extracted over the encoded expressions. (b) The training stages of the SchiNet.

do so in four steps. First, we detect the patient’s body at each frame using the Single Shot Detector (SSD) proposed in [30]. 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 SmileNet [31] to detect the bounding box of the face and whether it is smiling or not. Finally, we crop and scale the detected face to a fixed resolution of  $100 \times 100$  for further analysis. Note that no face-registration is applied to the extracted faces prior to expression 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 in more than 90% of the frames. By doing so, we retain the videos of 91 patients out of the total 110 that participated in the baseline session.

### 4.3 Low-Level Feature Extraction

In the second stage of the proposed method a DNN architecture is used to code the facial behaviour in terms of Facial Action Coding System (FACS) [33], that is, detect the activation of facial Action Units (AUs), and detect smiles, a specific facial behaviour, 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 [34], [35], [36], [37].

**Proposed AFEA Method.** Detection of facial AUs, i.e. detection of the activation of certain facial muscles, 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 Convolutional Neural Networks (CNNs) [21], [38], [39] since they have been shown to learn better and more general appearance features compared to hand-crafted ones [15]. Furthermore, networks trained on large datasets on surrogate tasks (e.g. object detection) have been shown to perform well for feature extraction on other tasks. Motivated by this, we refine VGG-16 [40] for the detection of 10 facial 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:

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

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

Since the occurrence of facial expressions is correlated, many works deal with facial expression detection as a multi-label classification problem [21], [41], [42]. In this work, we train a separate network for each expression, because the number of positive examples vary immensely from one expression 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 classes. Data balancing can alleviate this problem, however, this is hard in the case of multi-label problems and typically separate networks perform better.

In total, 11 facial expressions are analysed, ten of which are facial AUs detected using the AFEA method described above, and one is smile recognized using the SmileNet proposed in [31]. This results in an 11-dimensional feature vector for each frame where each dimension represents the probability of one of the detected expressions.

**AFEA for the NESS Dataset.** 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 behaviour 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 expression is calculated and subtracted from the activations of the expression in question.

#### 4.4 High-Level Feature Extraction

In section 4.3, we extracted frame level representations, i.e. at each frame  $t$  of the sequence we extracted a vector  $\mathbf{x}_t \in R^{11}$ , containing the probability of the occurrence of the 11 facial expressions. 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 expressions probabilities  $\mathbf{x} \in R^{11}$  using a weighted sum of  $K$  Gaussian distributions [43]. Clearly, the distribution is over the set of  $\mathbf{x}$  that are extracted over the whole training dataset, one  $\mathbf{x}$  for every frame of each sequence. In this context, each GMM component would represent a commonly occurring combination of facial expressions. The GMM is expressed as:

$$u_\lambda(\mathbf{x}) = \sum_{k=1}^K w_k u_k(\mathbf{x}), \quad (2)$$

where  $w_k$  is the weight component of the  $k$ -th Gaussian distribu-

tion  $u_k(\mathbf{x})$ .  $u_k(\mathbf{x})$  is defined as:

$$u_k(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)' \Sigma_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right). \quad (3)$$

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

$$\gamma_t(k) = \frac{w_k u_k(\mathbf{x}_t)}{\sum_l^K w_l u_l(\mathbf{x}_t)}. \quad (4)$$

In this work we follow [32], and implement the GMM as a neural network layer, that during training given a set of  $\mathbf{x}$  learns the parameters of the GMM and during testing given an  $\mathbf{x}$  produces  $K$  GMM posteriors  $\{\gamma_t(k), k = 1, \dots, K\}$  at its output (see Figure 1(a)). The GMM layer is first trained in unsupervised way using the Expectation-Maximization (EM) algorithm [44], 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  $\mathbf{x}_t$  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 [43], 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  $\Sigma_k$ . Formally:

$$\mathcal{G}_\lambda^X = \left( \mathcal{G}_{w_1}^X, \dots, \mathcal{G}_{w_K}^X, \mathcal{G}_{\mu_1}^{X'}, \dots, \mathcal{G}_{\mu_K}^{X'}, \mathcal{G}_{\sigma_1}^{X'}, \dots, \mathcal{G}_{\sigma_K}^{X'} \right)', \quad (5)$$

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

$$\mathcal{G}_{w_k}^X = (S_k^0 - T w_k) / \sqrt{w_k}, \quad (6)$$

$$\mathcal{G}_{\mu_k}^X = (\mathbf{S}_k^1 - \boldsymbol{\mu}_k S_k^0) / (\sqrt{w_k} \boldsymbol{\sigma}_k), \quad (7)$$

$$\mathcal{G}_{\sigma_k}^X = (S_k^2 - 2\boldsymbol{\mu}_k \mathbf{S}_k^1 + (\boldsymbol{\mu}_k^2 - \boldsymbol{\sigma}_k^2) S_k^0) / (\sqrt{2w_k} \boldsymbol{\sigma}_k^2), \quad (8)$$

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

$$S_k^0 = \sum_{t=1}^T \gamma_t(k), \quad (9)$$

$$\mathbf{S}_k^1 = \sum_{t=1}^T \gamma_t(k) \mathbf{x}_t, \quad (10)$$

and

$$\mathbf{S}_k^2 = \sum_{t=1}^T \gamma_t(k) \mathbf{x}_t^2, \quad (11)$$

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 [43], the extracted FV is normalized using both power normalization, and L2 normalization.

In [32], 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 expressions. The layer output or the FV has a length of  $K(2N + 1)$ , where  $K$  is the number of GMM components and  $N$  is the feature dimensionality, which in our case is the number of the detected expressions, that is 11. 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) to the FV dimensionality (368 for  $K = 16$  and  $N = 11$ ), 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.

#### 4.5 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 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:

$$C_d(p, t) = \frac{1}{V} \sum_{v=1}^V \frac{1}{W} \sum_{w=1}^W (p_{vw} - t_{vw})^2, \quad (12)$$

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.

## 5 EXPERIMENTS AND RESULTS

In this section, we report the performance of the developed architectures for facial expression analysis and symptom severity estimation. First, we test the performance of the proposed AFEA method and compare it with a state-of-the-art method in detecting facial expressions “in the wild”. Then, we measure the correlations between facial expressions and different symptoms of schizophrenia. Finally, we report the performance of the proposed SchiNet in estimating symptom severity and compare it to other works in the literature.

### 5.1 Classification of Facial Expressions

The interviews in the NESS dataset have a wide range of different camera poses and illuminations levels. Furthermore, patients tend to gaze down or away from the interviewer, or sometimes occlude the face with different hand gestures. In order to handle with such challenges it is imperative to train the facial expression analysis method, with datasets that contain such variations – in this work we relied on recent datasets collected “in the wild”.

**Datasets.** We use 4 datasets collected “in the wild”, for the detection of 10 facial expressions – Table 2 shows the used datasets, as well as the detected expressions. 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 greatly the robustness of our model to those conditions. For the EmotioNet dataset [45], only manually-annotated images in the validation set are used in the training and testing of the AFEA method. The EmotioNet consists of annotations for 12 expressions. Although we trained different networks for detecting the 12 expressions, only 7 expressions (shown in Table 2) show good performance when applied to the NESS dataset – those are selected for further analysis.

**Training Settings.** We split the datasets (CEW [48], CelebA [47], EmotioNet [45], ExpW [46]) into 75% for training, 10% for validation, and 15% for testing. Many of the detected expressions 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 [49]. The ExpW [46] 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.

The training set of each expression is augmented with random flipping, rotation, shifting, shearing, and zooming, in order to avoid over-fitting. We initialize the parameters of the expression recognition networks by the parameters of the VGG-16 and refine them using SGD with adaptive learning rate (RMSprop [50]), 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 expression, the batch size is set either to 64 or 128. The first column of Table 1 summarizes the parameter values used for training the AFEA part of the SchiNet.

**State-of-the-Art.** We compare our AFEA method with Bishay and Patras [21], a method that achieves state-of-the-art results on the BP4D dataset [51] for facial expression recognition. The comparison is two-fold. First we test how both methods perform on facial expression recognition on datasets collected “in the wild” (Sec. 5.1). Second, we apply both methods on the patients’ interviews for low-level feature extraction, retrain the methods for high-level feature extraction and symptom severity estimation using the corresponding features, and observe how the performance is affected (Sec. 5.3). In both comparisons, only the 8 expressions that are detected by both methods are used. For a fair comparison on the facial expression recognition part, we use only two of the several networks that we proposed and fused in [21], and more specifically, the 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). The reason for doing so is that the network adopted here is not trained on dynamic information. In the second comparison, i.e. testing how each AFEA method performs on symptom assessment, we compare with the full architecture. In what follows we will refer to the simplified static version of [21] as “Static [21]” and the full architecture as “Full [21]”.

**Results.** Accuracy and F1-score obtained by the proposed method on the 15% testing splits are shown in Table 2. We observe that the performance is highly dependent on the number

TABLE 1  
An overview of the settings used for the training of the SchiNet.

Training setting	AFEA	GMM-FV-FC1	FC2
Weight initialization	VGG-16	EM algorithm	Randomly
Update function	RMSProp	SGD with momentum	SGD with momentum
Learning rate	$10^{-4}$	0.005 (CAINS) / 0.001 (PANSS)	0.01
Momentum	-	0.9	0.9
Decay coefficient	0.7	-	-
Cost	Binary cross entropy	Mean square error	Mean square error
Batch size	64 / 128	Total no. of patients	Total no. of patients

TABLE 2

The classification results obtained by the frame-based components of the network proposed in [21] and by the proposed AFEA method over different facial expressions on the 15% testing splits of the EmotioNet [45], ExpW [46], CelebA [47], and CEW [48] datasets.

Facial Expressions		Inner Brow Raiser	Outer Brow Raiser	Brow Lowerer	Upper Lid Raiser	Cheek Raiser	Lip Corner Puller	Lips Part	Neutral Expression	Lid Tightener	Eyes Closed
		EmotioNet								ExpW	CelebA
Static [21]	Acc	0.679	0.669	0.752	-	0.806	0.823	0.708	-	0.670	0.617
	F1	0.166	0.130	0.354	-	0.544	0.792	0.697	-	0.268	0.422
	Avg	0.423	0.399	0.553	-	0.675	0.808	0.703	-	0.469	0.520
Proposed AFEA Method	Acc	<b>0.941</b>	<b>0.869</b>	<b>0.903</b>	0.857	<b>0.880</b>	<b>0.908</b>	<b>0.919</b>	0.731	<b>0.855</b>	<b>0.980</b>
	F1	<b>0.459</b>	<b>0.319</b>	<b>0.632</b>	0.304	<b>0.716</b>	<b>0.897</b>	<b>0.912</b>	0.718	<b>0.526</b>	<b>0.977</b>
	Avg	<b>0.700</b>	<b>0.594</b>	<b>0.767</b>	0.580	<b>0.798</b>	<b>0.902</b>	<b>0.915</b>	0.725	<b>0.691</b>	<b>0.979</b>

of training samples and the variance in expression-appearance. More specifically, expressions like lips part, and eyes closed have a high value for both F1-score and accuracy, due to the relatively large number of training examples as well as fewer differences in expression-appearance among subjects. On other expressions like brow lowerer, and lid tightener we obtain moderately good performance due to the large variance in expression-appearance among different people. Finally, we obtain low F1-score values for the outer brow raiser and the upper lid raiser as the EmotioNet dataset is highly imbalanced for those two classes.

In Table 2, we also show the performance of Static [21] on the testing splits – we observe that our method obtains better results in the 8 expressions. This considerable difference in performance is mainly due to two reasons. First, [21] is trained using facial images captured in a controlled environment, and with limited variation in head pose. Second, only 2 out of 8 deep networks in [21] are used in expressions detection. [21] shows that the full architecture achieves better performance than both single and combined networks. Figure 2 shows a qualitative comparison between the two AFEA methods on different facial images drawn from the testing splits. Each image has a face with an active facial expression. We can see that Static [21] performs well mainly for frontal or near-frontal faces, while the proposed method can detect expressions at several head poses and illumination levels. However, the proposed method fails when the expressions are subtle, or when the faces are captured under too dark or bright illumination conditions.

## 5.2 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 facial expressions, namely, the frequency of the occurrence of the expression in question, correlate with the different symptom scales of schizophrenia. We show that for several symptoms, high and significant correlations with facial expressions 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 [19], and CAINS [20]. 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 NEG scale, 7 form 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).

**Calculating Correlations.** We use the Spearman's correlation for measuring the association between the ground-truth symptom levels and activation frequency of each expression. 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 11 expressions, and then compute the activation frequency as follows:

$$f_i = \frac{N_i}{N_{total}}, \quad i \in \{1, 2, \dots, 11\} \quad (13)$$

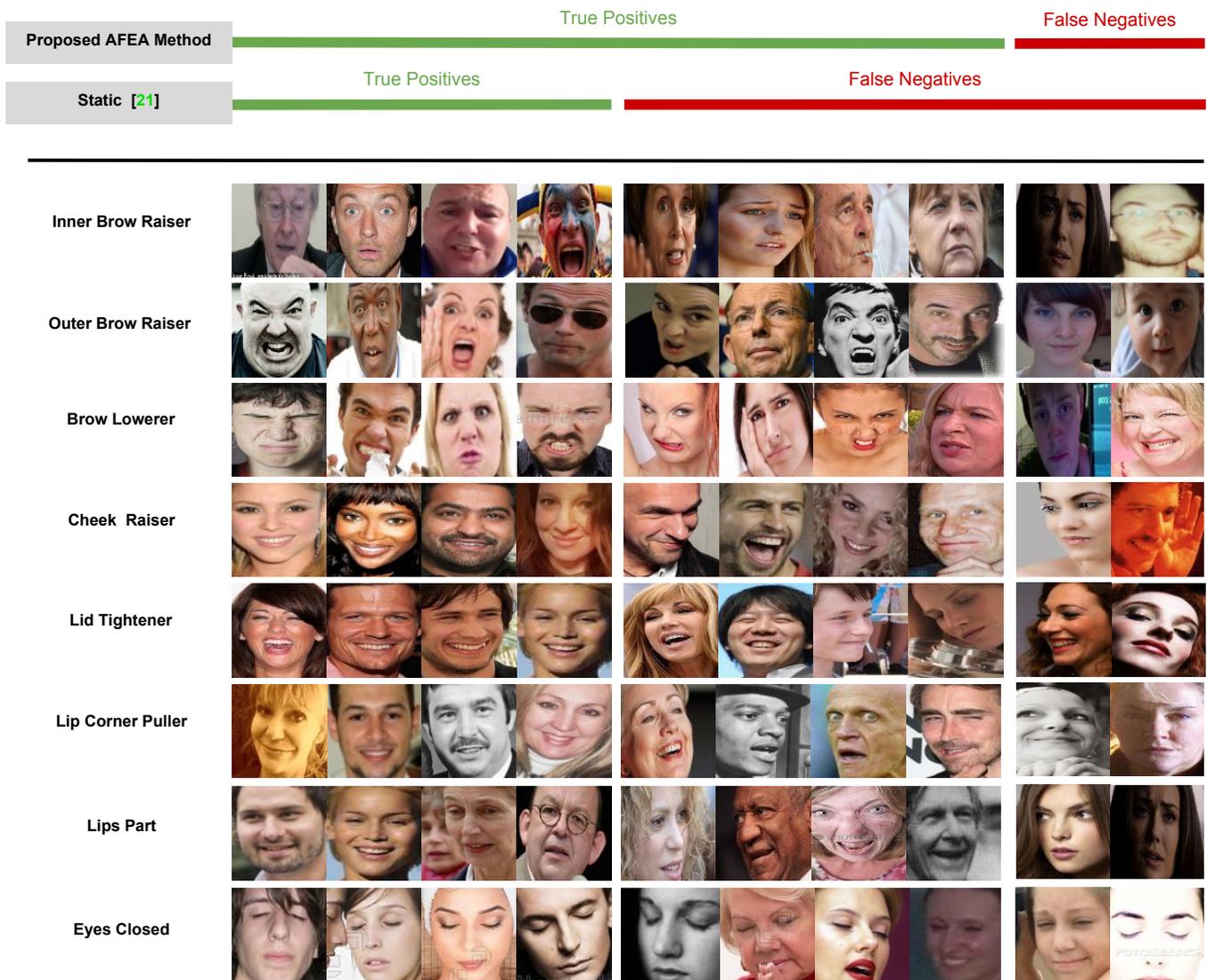


Fig. 2. Qualitative comparison between the proposed AFEA method, and Static [21] in expression analysis. Each row shows the positive examples of a certain facial expression. The true positives and false negatives achieved by each method are shown on the top part of the figure. The proposed method shows better performance in detecting expressions at several head poses and illumination conditions, compared to Static [21].

where  $N_i$  is the number of frames for which expression  $i$  is activated and  $N_{total}$  is the total number of video frames with a successful face detection.

The faces of some patients are occluded by a wearable item (e.g. thick eyeglasses) – this sometimes results in the related facial expression 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  $\sigma_i$  is the standard deviation over the frequencies of expression  $i$  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.** Table 3 and 4 show the correlations found between facial expressions on the one hand, and some symptoms in both of the CAINS and PANSS scales on the other. In CAINS (Table 3), significant associations are found between lips part, which is commonly activated during patients' speech, and symptoms like quantity of speech, vocal expression, and facial expression.

Similarly, in PANSS (Table 4), higher levels of symptoms like lack of spontaneity and flow of conversation, 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 facial expressions and the total score of the CAINS and PANSS scales. For instance, CAINS-EXP scale has significant associations with many facial expressions e.g. neutral expression, cheek raiser and lips part. As the NESS dataset contains patients with a relatively high level of negative symptoms and lower levels of positive and general psychopathology symptoms, more significant correlations are found with negative symptoms.

TABLE 3  
Correlations found between facial expressions and the **CAINS** symptoms.

Symptoms	Facial Expressions							
	Neutral Expression	Cheek Raiser	Lid Tightener	Lip Corner Puller	Lips Part	Smiling	Eyes Closed	
<b>EXP</b> - Facial Expression	0.45**	-0.43**	-	-0.4**	-0.33*	-0.42**	-	
<b>EXP</b> - Vocal Expression	0.35**	-0.38**	-0.34*	-	-0.41**	-	-	
<b>EXP</b> - Expressive Gestures	-	-0.32*	-	-	-0.43**	-	-	
<b>EXP</b> - Quantity of Speech	0.38**	-	-	-	-0.41**	-	-	
<b>MAP</b> - Motivation for Recreational Activities	-	-	-	-	-	-	-0.47**	
<b>MAP</b> - Frequency of Pleasurable Recreational Activities - Past Week	-	-	-	-	-	-	-0.35**	
<b>EXP</b> - Total Score	0.42**	-0.41**	-	-	-0.46**	-0.29*	-	

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

TABLE 4  
Correlations found between facial expressions and the **PANSS** symptoms.

Symptoms	Facial Expressions							
	Neutral Expression	Inner Brow Raiser	Outer Brow Raiser	Cheek Raiser	Lid Tightener	Lips Part	Smiling	
<b>NEG</b> - Flat Affect	0.28*	-	-	0.33**	-	-0.37**	-0.29*	
<b>NEG</b> - Poor Rapport	-	-	-	-0.36**	-	-0.34*	-0.28*	
<b>NEG</b> - Lack of Spontaneity and Flow of Conversation	0.32*	-	-	-	-	-0.31*	-	
<b>POS</b> - Suspiciousness/Persecution	-	-	0.36**	-	-	-	-	
<b>GEN</b> - Somatic Concern	-	0.29*	-	-	0.33*	-	-	
<b>GEN</b> - Anxiety	-	-	0.29*	-	-	-	-	
<b>NEG</b> - Total Score	-	-	-	-	-	-0.30*	-0.30*	
<b>POS</b> - Total Score	-	0.31*	0.30*	-	-	0.29*	-	
<b>GEN</b> - Total Score	-	-	-	-	0.37**	-	-	

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

### 5.3 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 [18], 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. Based on that, we focus in this work on assessing the highly correlated negative symptoms in both the CAINS and PANSS interviews through 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 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 [52], [53] for implementing the GMM, FV, and FC layers. The number of GMM components ( $K$ ) is set to 16, giving for each video a FV with length 368. Following [43], 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 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

TABLE 5  
Comparison between the proposed AFEA method, and Full [21] on estimating CAINS-EXP symptoms.

	Full [21]			Proposed AFEA Method		
	PCC	MAE	RMSE	PCC	MAE	RMSE
<b>EXP - Facial Expression</b>	0.37	0.80	1.07	<b>0.42</b>	<b>0.74</b>	<b>0.99</b>
<b>EXP - Vocal Expression</b>	0.25	0.93	1.22	<b>0.30</b>	<b>0.75</b>	<b>1.13</b>
<b>EXP - Expressive Gestures</b>	0.04	1.07	1.37	<b>0.34</b>	<b>0.99</b>	<b>1.19</b>
<b>EXP - Quantity of Speech</b>	<b>0.42</b>	1.07	1.37	0.39	<b>0.91</b>	<b>1.22</b>
<b>EXP - Total Score</b>	0.29	3.06	3.80	<b>0.42</b>	<b>2.90</b>	<b>3.61</b>

TABLE 6  
Comparison between the proposed AFEA method, and Full [21] on estimating PANSS-NEG symptoms.

	Full [21]			Proposed AFEA Method		
	PCC	MAE	RMSE	PCC	MAE	RMSE
<b>NEG - Flat Affect</b>	0.21	0.97	1.31	<b>0.32</b>	<b>0.96</b>	<b>1.27</b>
<b>NEG - Poor Rapport</b>	0.25	0.91	1.22	<b>0.41</b>	<b>0.75</b>	<b>1.13</b>
<b>NEG - Lack of Spontaneity and Flow of Conversation</b>	0.13	<b>1.28</b>	1.60	<b>0.24</b>	<b>1.28</b>	<b>1.54</b>
<b>NEG - Total Score</b>	0.08	4.00	4.88	<b>0.40</b>	<b>3.35</b>	<b>4.27</b>

$lr = 0.01$ . 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 (LOOCV) is used for validating and testing our architecture. The second and the third column of Table 1 summarize the parameter values used for training the GMM-FV-FC1 and FC2 layers of the SchiNet, respectively.

**Performance Measures.** Three measures are used for reporting the performance of the symptom severity estimation using as ground truth the psychiatrists’ assessments. Following [12], [14], 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 AFEA Architecture.** In this experiment we evaluate how our AFEA method performs on symptom severity estimation compared to [21]. In order to do so, the full architecture in [21] (Full [21]) is applied for analysing the videos of all of the 110 baseline patients in the NESS dataset. Out of the 110, we considered only 74 patients for whom we can successfully detect faces and landmarks in more than 90% of the frames of their videos. Two output neurons are used to predict, respectively, the presence and absence of each expression in Full [21] – 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. The mean over each expression is subtracted (normalization step), and then the expressions probabilities are used as input to the GMM layer.

As the number of patients and facial expressions analysed by the two AFEA methods vary, only the ones that are common in both methods are used in the comparison. Based on that, 69 patients and 8 common expressions (shown in Table 2) are selected for symptom estimation. The number of GMM components is set

to 12 in this case, as fewer patients and expressions are analysed. Table 5 and 6 show the estimation results of the CAINS and PANSS symptoms, respectively, using both the proposed AFEA method, and the Full [21]. From the comparison, we can see that the proposed AFEA method leads to better symptom estimation in all the estimated symptoms. This illustrates the positive impact of training the AFEA method using data captured “in the wild”.

**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.* [12], [14]. 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 [14], 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 [12], [14] estimate specific symptoms of schizophrenia and not the total score, we discard the total CAINS-EXP and PANSS-NEG scores in the comparison. The 91 patients analysed by the proposed AFEA method are used in the comparison. Table 7 and 8 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 hand-crafted ones.

For estimating the total PANSS-NEG score, the proposed architecture uses the 3 out of 7 PANSS-NEG symptoms that are highly correlated with facial expressions, as input to the FC2 layer. In Table 9, 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, then estimating the total score at FC2 layer. In both

TABLE 7  
Performance of the SchiNet as well as other state-of-the-art methods on the **CAINS-EXP** symptoms.

	Tron <i>et al.</i> [12]			Tron <i>et al.</i> [14]			SchiNet		
	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE
<b>EXP - Facial Expression</b>	0.37	0.80	1.03	0.36	0.75	1.07	<b>0.46</b>	<b>0.66</b>	<b>0.93</b>
<b>EXP - Vocal Expression</b>	0.23	0.87	1.23	0.26	0.86	1.22	<b>0.27</b>	<b>0.77</b>	<b>1.10</b>
<b>EXP - Expressive Gestures</b>	0.36	<b>0.85</b>	1.19	<b>0.38</b>	0.91	1.22	0.36	0.90	<b>1.15</b>
<b>EXP - Quantity of Speech</b>	0.27	1.09	1.43	0.25	1.02	1.36	<b>0.30</b>	<b>0.98</b>	<b>1.30</b>
<b>EXP - Total Score</b>	-	-	-	-	-	-	<b>0.45</b>	<b>2.67</b>	<b>3.34</b>

TABLE 8  
Performance of the SchiNet as well as other state-of-the-art methods on the **PANSS-NEG** symptoms.

	Tron <i>et al.</i> [12]			Tron <i>et al.</i> [14]			SchiNet		
	PCC	MAE	RMSE	PCC	MAE	RMSE	PCC	MAE	RMSE
<b>NEG - Flat Affect</b>	0.37	0.90	1.28	0.11	0.99	1.36	<b>0.42</b>	<b>0.84</b>	<b>1.18</b>
<b>NEG - Poor Rapport</b>	0.20	0.98	1.31	0.15	1.01	1.26	<b>0.27</b>	<b>0.85</b>	<b>1.20</b>
<b>NEG - Lack of Spontaneity and Flow of Conversation</b>	0.13	1.37	1.69	0.09	1.32	1.62	<b>0.25</b>	<b>1.25</b>	<b>1.51</b>
<b>NEG - Total Score</b>	-	-	-	-	-	-	<b>0.29</b>	<b>3.30</b>	<b>4.17</b>

TABLE 9  
The estimation results of the total **PANSS-NEG** score obtained in different settings.

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

cases, the results were worse. In the first case a possible reason is that NEG symptoms have more significant correlations with expressions 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 facial expressions, making the training of the FC2 layer worse.

## 6 CONCLUSIONS AND FUTURE WORK

Our work aims to develop an automatic tool that is capable of quantifying patient behaviour, and then using it 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. Analysing interviews of patients at a wide variety of poses and illumination conditions led us to implement an AFEA method that is trained using data collected “in the wild”, that is outside laboratory conditions. Then, patients’ facial expressions are detected and 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 AFEA “in the wild” performs better on symptom estimation than another state-of-the-art method [21]

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 – 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 is a feasible and reliable method 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 [29] (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 plan in our future work to improve the performance of the AFEA method, by moving from static to temporal analysis “in the wild”. Moreover, we will extend the behaviour analysis to include body gestures and vocal expressions (besides facial expressions). Finally, we will explore how the symptom estimation can be improved using a personalized model.

## ACKNOWLEDGMENTS

The work of Mina Bishay is a part of the Newton-Mosharafa PhD scholarship, which is jointly funded by the Egyptian Ministry of Higher Education and the British Council.

## REFERENCES

- [1] F. Trémeau, "A review of emotion deficits in schizophrenia," *Dialogues in clinical neuroscience*, vol. 8, no. 1, p. 59, 2006.
- [2] M. K. Mandal, R. Pandey, and A. B. Prasad, "Facial expressions of emotions and schizophrenia: A review." *Schizophrenia bulletin*, vol. 24, no. 3, p. 399, 1998.
- [3] 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*, vol. 64, no. 3, pp. 375–378, 1998.
- [4] D. Murphy and J. Cutting, "Prosodic comprehension and expression in schizophrenia." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 53, no. 9, pp. 727–730, 1990.
- [5] M. Brüne, C. Sonntag, M. Abdel-Hamid, C. Lehmkämpfer, G. Juckel, and A. Troisi, "Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders," *The Journal of nervous and mental disease*, vol. 196, no. 4, pp. 282–288, 2008.
- [6] A. Troisi, G. Spalletta, and A. Pasini, "Non-verbal behaviour deficits in schizophrenia: an ethological study of drug-free patients," *Acta Psychiatrica Scandinavica*, vol. 97, no. 2, pp. 109–115, 1998.
- [7] 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*, vol. 202, no. 1, pp. 47–54, 2014.
- [8] 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*, vol. 43, no. 4, pp. 216–222, 2010.
- [9] M. Lavelle, P. G. Healey, and R. McCabe, "Is nonverbal communication disrupted in interactions involving patients with schizophrenia?" *Schizophrenia bulletin*, vol. 39, no. 5, pp. 1150–1158, 2012.
- [10] E. Worswick, S. Dimic, C. Wildgrube, and S. Priebe, "Negative symptoms and avoidance of social interaction: A study of non-verbal behaviour," *Psychopathology*, 2017.
- [11] 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*, vol. 163, no. 2, pp. 350–361, 2007.
- [12] 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*. Springer, 2015, pp. 72–81.
- [13] 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*, vol. 168, no. 1, pp. 224–238, 2008.
- [14] T. Tron, A. Peled, A. Grinsphoon, and D. Weinshall, "Facial expressions and flat affect in schizophrenia, automatic analysis from depth camera data," in *Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on*. IEEE, 2016, pp. 220–223.
- [15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [16] 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*. IEEE, 2011, pp. 3361–3368.
- [17] 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*. ACM, 2015, pp. 1263–1266.
- [18] 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*, vol. 45, no. 8, pp. 1613–1627, 2015.
- [19] S. R. Kay, A. Flszbein, and L. A. Opfer, "The positive and negative syndrome scale (panss) for schizophrenia." *Schizophrenia bulletin*, vol. 13, no. 2, p. 261, 1987.
- [20] 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*, vol. 132, no. 2, pp. 140–145, 2011.
- [21] 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*. IEEE, 2017, pp. 681–688.
- [22] P. Davison, C. Frith, P. Harrison-Read, and E. Johnstone, "Facial and other non-verbal communicative behaviour in chronic schizophrenia," *Psychological medicine*, vol. 26, no. 04, pp. 707–713, 1996.
- [23] 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*. IEEE, 2007, pp. 1–8.
- [24] 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*, vol. 200, no. 2, pp. 237–256, 2011.
- [25] 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*, vol. 2014, 2014.
- [26] 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*. IEEE, 2016, pp. 2427–2430.
- [27] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 41, 2013.
- [28] N. C. Andreasen, "Scale for the assessment of negative symptoms (sans)." *The British Journal of Psychiatry*, 1989.
- [29] S. Priebe, M. Savill, T. Wykes, R. Bentall, U. Reininghaus, C. Lauber, S. Bremner, S. Eldridge, and F. Röhrich, "Effectiveness of group body psychotherapy for negative symptoms of schizophrenia: multicentre randomised controlled trial," *The British Journal of Psychiatry*, pp. bjp–bp, 2016.
- [30] 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*. Springer, 2016, pp. 21–37.
- [31] 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*, 2017, pp. 1581–1589.
- [32] P. Palasek and I. Patras, "Discriminative convolutional fisher vector network for action recognition," *arXiv preprint arXiv:1707.06119*, 2017.
- [33] 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.
- [34] 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*. ACM, 2006, pp. 162–170.
- [35] 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*. ACM, 2014, pp. 65–72.
- [36] J. F. Cohn and F. De la Torre, "Automated face analysis for affective," *The Oxford handbook of affective computing*, p. 131, 2014.
- [37] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, 2017.
- [38] 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)*. IEEE, 2016, pp. 1–8.
- [39] 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*, 2016, pp. 3391–3399.
- [40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [41] 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*. IEEE, 2015, pp. 609–615.
- [42] A. Gudi, H. E. Tasli, T. M. den Uyl, and A. Maroulis, "Deep learning based face action unit occurrence and intensity estimation," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6. IEEE, 2015, pp. 1–5.
- [43] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.
- [44] 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)*, pp. 1–38, 1977.

- [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*, 2016, pp. 5562–5570.
- [46] C. C. L. Zhanpeng Zhang, Ping Luo and X. Tang, "From facial expression recognition to interpersonal relation prediction," in *arXiv:1609.06426v2*, 2016.
- [47] 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*, 2015, pp. 3730–3738.
- [48] 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*, vol. 47, no. 9, pp. 2825–2838, 2014.
- [49] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 853–867.
- [50] 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*, vol. 4, no. 2, pp. 26–31, 2012.
- [51] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.
- [52] 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.
- [53] 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*, vol. 3, 2015.



**Stefan Priebe** graduated in Psychology and Medicine, and qualified as Neurologist, Psychiatrist and Psychotherapist in Germany. Since 1997 he has been Professor for Social and Community Psychiatry at Queen Mary, University of London (QMUL). He is also Director of the WHO Collaborating Centre for Mental Health Service Development, Director of the NIHR Global Health Research Group for Developing Psycho-Social Interventions, Deputy Director of a registered Clinical Trials Unit, Research Director of the Institute for Population Health Sciences (all at QMUL) and R&D Director of East London NHS Foundation Trust.

He heads a research group in East London which runs several programmes focusing on understanding, modifying and utilising social interactions to reduce mental distress.



**Mina Bishay** received the BSc and MSc degrees (with honors) in electrical engineering (Electronics and Communications section) from Assiut University, Egypt, in 2010 and 2014, respectively. During his MSc thesis, he focused on improving the performance of fingerprint image segmentation and enhancement. He received the Best Student Paper Award in IEEE National Radio Science Conference (NRSC) in 2014. He is now working towards the PhD degree at the School of Electronic Engineering and Computer

Science, Queen Mary University of London, UK. His research interests include: automatic behaviour analysis, affective computing and deep learning.



**Ioannis Patras** received the BSc and MSc degrees in computer science from the Computer Science Department, University of Crete, Heraklion, Greece, in 1994 and 1997, respectively, and the PhD degree from the Department of Electrical Engineering, Delft University of Technology, Delft (TU Delft), The Netherlands, in 2001. He is a Professor in Computer Vision and Human Sensing in the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK. His current research

interests are in the area of Computer Vision, Machine Learning and Affective Computing, with emphasis on the analysis of visual data depicting humans and their activities. He is an Associate Editor of the *Image and Vision Computing Journal*, *Pattern Recognition*, and *Computer Vision and Image Understanding*.



**Petar Palasek** received the BSc and MSc degrees in computer science from the Faculty of Electrical Engineering and Computing, University of Zagreb in 2010 and 2012 respectively, and the PhD degree from the School of Electronic Engineering and Computer Science, Queen Mary University of London in 2017. During his PhD, his research focused on studying deep learning architectures for human action recognition in videos, and it also included the problems of 3D human reconstruction and human

body pose estimation. His research interests include computer vision, machine learning and neural networks. He is currently working as a research scientist at MindVisionLabs LTD, a London based startup.