CAN AUTOMATIC FACIAL EXPRESSION ANALYSIS BE USED FOR TREATMENT OUTCOME ESTIMATION IN SCHIZOPHRENIA?

Mina Bishay, Stefan Priebe and Ioannis Patras

Queen Mary University of London, UK

ABSTRACT

Negative symptoms of schizophrenia include expressive deficits that are marked by a reduction in patients' behaviour. Analysing automatically non-verbal behaviour and exploiting the results for estimating symptom severity has drawn attention recently. However, those approaches are not accurate enough to be used for monitoring the changes in patient's symptom level during treatment interventions (i.e. the treatment outcome). In this paper, we propose a method that directly addresses the problem of Treatment Outcome Estimation (TOE) in schizophrenia – more specifically, is 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. The proposed architecture builds on Recurrent Neural Networks (RNNs) that learn differences in the patient behaviour before and after treatment. We validate the method in videotaped interviews for symptom assessment for 74 patients. Experimental results show that the proposed architecture achieves promising results for TOE in two different symptom assessment scales.

Index Terms— Facial expression analysis, schizophrenia, negative symptoms, treatment outcome estimation

1. INTRODUCTION

Schizophrenia is a severe mental health condition affecting the way a patient thinks, feels, and behaves. Negative symptoms of schizophrenia are persistent [1], and have a greater effect on patients' quality of life in comparison to other symptoms [2]. These symptoms include impairments in the expression of emotion and speech (e.g. flat affect, impoverished speech) [3], which are observed through a diminution in the patients' non-verbal behaviour during social/clinical interventions [4]. For symptom assessment, non-verbal behaviour is rated subjectively during clinical interviews.

Recently, automatic facial expression analysis has been used for studying differences in behaviour between patients with schizophrenia and healthy controls [5, 6], for diagnosis and for the estimation of the severity of the symptoms [7, 8]. However, the proposed methods for symptom estimation are not suitable for determining whether the symptoms have improved or not. The reason for that is that often the change in negative symptoms is small [9], and falls within the error margin of these methods.

In this paper, we propose a deep neural network architecture that directly addresses the problem of treatment outcome estimation in schizophrenia. The architecture jointly analyses 2 video interviews of a patient, one recorded before and one recorded after the treatment, and gives as output the treatment outcome, that is a binary label that encodes whether the symptom has improved or not. The proposed architecture consists of three main stages. Firstly, we detect automatically the patient's behaviour (facial expressions) in the pair of videos. Then, we use a Recurrent Neural Network (RNN) for learning the local differences/changes in the patient behaviour over short concatenated clips from both videos. Finally, another RNN uses the clip-level features for learning global (i.e. patient-level) features, and outputs the treatment outcome.

The architecture is trained in a patient-independent manner on a dataset of 74 patients with 148 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) [10], and Positive and Negative Syndrome Scale (PANSS) [11]. Our architecture delivers promising results.

2. RELATED WORK

In psychiatry, a lot of research focused on studying the non-verbal behaviour of patients with schizophrenia [12, 13]. In these works, the non-verbal behaviour was manually annotated by human raters. Manual annotation is a rigorous, time-consuming process. Moreover, non-verbal behaviour is rated subjectively during clinical assessments. For these reasons, several works investigate on the utilisation of Automatic Facial Expression Analysis (AFEA) for a) studying patient behaviour, and b) diagnosing schizophrenia.

Studying patient behaviour. Many works [5, 6, 14, 15] focused on comparing patients' behaviour to that of healthy

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.

controls, by extracting and comparing behaviour-related features of patients and controls. Furthermore, associations between these features and different symptoms were examined. Different AFEA methods and features were used. In [5], the probabilities of four emotion categories were detected in single frames and then their average over the whole video was calculated. In [6], Wang *et al.* used temporal facial information for detecting 4 emotional and the neutral expressions, then features like their occurrence frequencies were calculated. In [14, 15], 15 Action Units (AUs) were detected per frame - in [14], features like the frequency of some single and combined AUs were extracted, while in [15] Information Theory measures were used for assessing the ambiguity and distinctiveness of participants' facial expressions. In all of those studies, patients and controls were recorded in a bright room while listening to emotional situations about their life.

Diagnosis of schizophrenia. In [7, 8, 16], AFEA was used for the diagnosis or/and symptom severity estimation in schizophrenia. In [7, 8], a commercial 3D facial analysis tool was applied for extracting 23 AUs. Then, in [7] features related to each AU intensity and dynamics were extracted, while in [8], Tron *et al.* used clustering analysis over all AUs for extracting 3 features related to expression flatness. These features were then used for training a two-step SVM-based algorithm, first for separation between patients and healthy controls, and then for the estimation of some PANSS symptoms. In [7, 8], the subjects were recorded while answering emotionally evocative questions. In [16], Bishay *et al.* proposed to turn the analysis from controlled contexts and environments to settings that are similar to the ones found in clinics. Furthermore, a deep architecture (called SchiNet) was proposed for analysing patients' facial expressions in the wild, as well as estimating symptom severity by extracting deep statistical features using Gaussian mixture model and Fisher Vector layers.

To the best of our knowledge, no work addresses directly the problem of treatment outcome estimation. Some of the works reviewed above could be used for this purpose, by estimating the symptom level before and after 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.

3. THE DATA OF SCHIZOPHRENIA

In this work we use recordings and symptom annotations from the "NESS" trial [17], that was collected for evaluating body psychotherapy as a treatment for negative symptoms of schizophrenia. The participants in the NESS trial were recruited from mental health centres at 4 different locations across the UK; East London, South London, Manchester, and Liverpool. In total 275 participants were included in this 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). Assessment interviews took between 40-120 minutes for each patient. Several scales were used for measuring the outcome of the treatment such as PANSS [11] including negative, positive and general psychopathology symptoms, and CAINS [10] 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 2 out of the 3 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 treatment). Most of the videos were recorded at a frame rate of 25 fps 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 [17].

4. PROPOSED METHOD

In this section we present the proposed method for treatment outcome estimation. Figure 1 shows an overview of the method. 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 (RNNs).

4.1. 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, and the long video into N equally-spaced clips, as shown in Figure 1. The sliced clips are then down-sampled by a factor of 3 to reduce redundancy. A pair of clips (one from each video) is then passed to the next processing steps.

For each frame in the paired clips, we detect the patient's body using [18], and then within the body region we extract the face bounding-box and the smiling behaviour using the SmileNet proposed in [19]. The extracted face is then scaled

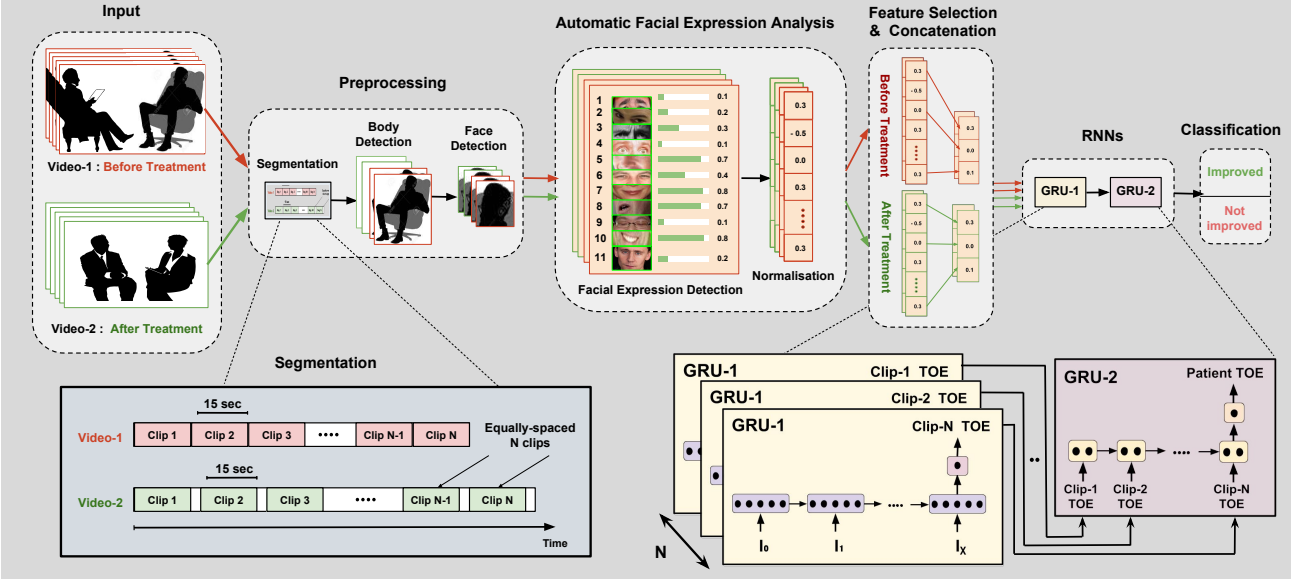

Fig. 1. The proposed architecture for treatment outcome estimation in schizophrenia.

to a fixed resolution of 100×100 and passed to the facial expression analysis architecture.

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 the 88 included from the NESS trial. Note that video-1 and video-2 of each patient should meet this condition in order for the patient to be included.

4.2. Automatic Facial Expression Analysis

Automatic Facial Expression Analysis (AFEA) is an active area of research that only recently has been moving from posed expressions and controlled environments, into spontaneous expressions in the wild (i.e. in real world conditions) [20]. In this work we use the AFEA methods proposed in [16] and in [19] – while the specific choice of the AFEA method is not crucial, it is important that it can deal with illumination, pose and scale variations as those are common in the data that we have. For each frame in the clips, we get an 11-dimensional feature vector, 10 dimensions corresponding to the probabilities of the presence of 10 facial action units (i.e. facial expressions) [16], and one corresponding to the probability of the presence of a smile [19].

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 expression is calculated and subtracted from the expression probabilities in the whole video (normalisation step).

4.3. Feature Selection

Feature selection is a crucial processing step in our architecture, as a relatively small number of patients are available for training. Sequential forward feature selection is used for selecting the most relevant expressions to the estimated symptoms. The selected expressions from video-1 (before treatment) and video-2 (after treatment) are concatenated at each time step and used as input to the RNNs. Note that the same expressions are selected in both videos.

4.4. Stacked RNNs for Treatment Outcome Estimation

Our architecture consists of two stacked RNNs (GRU-1, GRU-2), shown in Figure 1, and takes as input pairs of clips 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 (TOE) as a binary classification problem using RNNs. In this work, we adopt a GRU [21] 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 [22].

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 expressions probabilities in the pairs of clips for clip-level TOE. During training, GRU-1 is supervised by the patient-level labels. 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 fea-

Table 1. Performance of our architecture as well as other SSE methods on TOE for the **CAINS** expression symptoms.

Symptom	**Facial Expression**			**Vocal Expression**			**Expressive Gestures**			**Quantity of Speech**		
	F1	**Acc**	**Avg**	**F1**	**Acc**	**Avg**	**F1**	**Acc**	**Avg**	**F1**	**Acc**	**Avg**
Tron *et al.* [7]	0.18	0.62	0.40	0.27	0.59	0.43	0.29	0.61	0.45	0.21	0.60	0.41
Tron *et al.* [8]	0.14	**0.67**	0.41	0.31	**0.66**	0.49	0.22	0.67	0.45	0.07	0.62	0.35
SchiNet [16]	0.20	0.62	0.41	0.24	0.63	0.44	0.35	0.65	0.5	0.27	0.64	0.46
Ours	**0.42**	0.66	**0.54**	**0.43**	0.64	**0.54**	**0.37**	**0.70**	**0.54**	**0.33**	**0.68**	**0.51**

Table 2. Performance of our architecture as well as other SSE methods on TOE for the **PANSS** negative symptoms.

Symptom	**Flat Affect**			**Poor Rapport**			**Lack of Spontaneity**		
	F1	**Acc**	**Avg**	**F1**	**Acc**	**Avg**	**F1**	**Acc**	**Avg**
Tron *et al.* [7]	0.35	0.62	0.49	0.25	0.64	0.45	0.33	0.64	0.49
Tron *et al.* [8]	0.28	0.62	0.45	0.12	**0.71**	0.42	0.22	0.62	0.42
SchiNet [16]	0.31	0.64	0.48	0.20	0.67	0.44	0.21	0.60	0.41
Ours	**0.40**	**0.66**	**0.53**	**0.30**	0.68	**0.49**	**0.46**	**0.71**	**0.59**

tures/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. EXPERIMENTS AND RESULTS

Training Settings. We use 74 pairs of videos (one for each of the 74 patients) for training and testing our architecture using a Leave-One-Subject-Out (LOSO) protocol. More specifically, for each fold in the LOSO, 67 patients are used for training, 6 patients for validation, and 1 patient for testing. Training is done 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 RNNs using stochastic gradient descent with adaptive learning rate (RMSprop [23]), 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 batches) for GRU-2. 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 change the ground truth label accordingly to get an extra data sample.

We train the proposed architecture 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 [10], and 3 symptoms (flat affect, poor rapport, lack of spontaneity and flow of conversation) in the PANSS interview [11], are estimated. Note that a separate network is trained for each symptom.

Symptom Severity Estimation (SSE). In order to test how SSE methods perform on Treatment Outcome Estimation (TOE), 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 compare our architecture with three methods that have been proposed for SSE in schizophrenia [7, 8, 16] – for a fair comparison, we have re-implemented and re-trained those methods using the 74 patients. We used the probabilities of the 11 detected expressions 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 [8, 16], we tried different number of clusters or Gaussian components, and report the results of the best performing ones (12 clusters for [8] and 32 Gaussian components for [16]).

Results. In order to evaluate the performance of the proposed architecture, we report both the accuracy and F1 score, as well as the average over them. Table 1 and 2 summarise the performance of the SSE methods and the proposed architecture on TOE for the CAINS and PANSS symptoms, respectively. The SSE methods show relatively low performance when applied for TOE. The reason for that is that often the change in negative symptoms during treatment is small [9], and falls within their error margin. On average, the proposed architecture outperforms the SSE methods in all the CAINS and PANSS symptoms.

6. CONCLUSION

In this paper we propose an architecture that addresses directly the problem of TOE in schizophrenia. The architecture consists of RNNs that use facial expressions for learning local and global features over videos recorded before and after treatment. Symptom assessment interviews recorded in settings similar to real clinical ones are used in our analysis. Different expression-related negative symptoms from the PANSS and CAINS scales are estimated. The proposed architecture shows better performance in TOE, in comparison to other methods proposed for 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.

7. REFERENCES

[1] Hans-Jürgen Möller, "Clinical evaluation of negative symptoms in schizophrenia," *European Psychiatry*, vol. 22, no. 6, pp. 380–386, 2007.

[2] Beng-Choon Ho, Peg Nopoulos, Michael Flaum, Stephan Arndt, and Nancy C Andreasen, "Two-year outcome in first-episode schizophrenia: predictive value of symptoms for quality of life," *Focus*, vol. 155, no. 1, pp. 1196–137, 2004.

[3] Fabien Trémeau, "A review of emotion deficits in schizophrenia," *Dialogues in clinical neuroscience*, vol. 8, no. 1, pp. 59, 2006.

[4] Manas K Mandal, Rakesh Pandey, and Akhouri B Prasad, "Facial expressions of emotions and schizophrenia: A review.," *Schizophrenia bulletin*, vol. 24, no. 3, pp. 399, 1998.

[5] Christopher Alvino, Christian Kohler, Frederick Barrett, Raquel E Gur, Ruben C Gur, and Ragini Verma, "Computerized measurement of facial expression of emotions in schizophrenia," *Journal of neuroscience methods*, vol. 163, no. 2, pp. 350–361, 2007.

[6] Peng Wang, Frederick Barrett, Elizabeth Martin, Marina Milonova, et al., "Automated video-based facial expression analysis of neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 168, no. 1, pp. 224–238, 2008.

[7] Talia Tron, Abraham Peled, Alexander Grinsphoon, and Daphna 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.

[8] Talia Tron, Abraham Peled, Alexander Grinsphoon, and Daphna 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.

[9] Mark 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.

[10] William P Horan, Ann M Kring, Raquel E Gur, Steven P Reise, and Jack 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.

[11] Stanley R Kay, Abraham Flszbein, and Lewis A Opfer, "The positive and negative syndrome scale (PANSS) for schizophrenia.," *Schizophrenia bulletin*, vol. 13, no. 2, pp. 261, 1987.

[12] Alfonso 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.

[13] Elizabeth Worswick, Sara Dimic, Christiane Wildgrube, and Stefan Priebe, "Negative Symptoms and Avoidance of Social Interaction: A Study of Non-Verbal Behaviour," *Psychopathology*, 2017.

[14] Jihun Hamm, Christian G Kohler, Ruben C Gur, and Ragini Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, no. 2, pp. 237–256, 2011.

[15] Jihun Hamm, Amy Pinkham, Ruben C Gur, Ragini Verma, and Christian G Kohler, "Dimensional information-theoretic measurement of facial emotion expressions in schizophrenia," *Schizophrenia research and treatment*, vol. 2014, 2014.

[16] Mina Bishay, Petar Palasek, Stefan Priebe, and Ioannis Patras, "SchiNet: Automatic Estimation of Symptoms of Schizophrenia from Facial Behaviour Analysis," *arXiv preprint arXiv:1808.02531*, 2018.

[17] S Priebe, M Savill, T Wykes, RP 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*, pp. bjp–bp, 2016.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg, "SSD: Single Shot Multibox Detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[19] Youngkyoon Jang, Hatice Gunes, and Ioannis 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.

[20] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, 2017.

[21] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] Tijmen Tieleman and Geoffrey 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.