Spectral Representation of Behaviour Primitives for Depression Analysis

Siyang Song, Shashank Jaiswal, Linlin Shen, and Michel Valstar

Abstract—Depression is a serious mental disorder affecting millions of people all over the world. Traditional clinical diagnosis methods are subjective, complicated and require extensive participation of clinicians. Recent advances in automatic depression analysis systems promise a future where these shortcomings are addressed by objective, repeatable, and readily available diagnostic tools to aid health professionals in their work. Yet there remain a number of barriers to the development of such tools. One barrier is that existing automatic depression analysis algorithms base their predictions on very brief sequential segments, sometimes as little as one frame. Another barrier is that existing methods do not take into account what the context of the measured behaviour is. In this paper, we extract multi-scale video-level features for video-based automatic depression analysis. We propose to use automatically detected human behaviour primitives as the low-dimensional descriptor for each frame. We also propose two novel spectral representations, i.e. spectral heatmaps and spectral vectors, to represent video-level multi-scale temporal dynamics of expressive behaviour. Constructed spectral representations are fed to Convolution Neural Networks (CNNs) and Artificial Neural Networks (ANNs) for depression analysis. We conducted experiments on the AVEC 2013 and AVEC 2014 benchmark datasets to investigate the influence of interview tasks on depression analysis. In addition to achieving state of the art accuracy in severity of depression estimation, we show that the task conducted by the user matters, that fusion of a combination of tasks reaches highest accuracy, and that longer tasks are more informative than shorter tasks, up to a point.

Index Terms—Automatic depression analysis, Fourier Transform, Spectral representation, Time-frequency analysis, Convolution Neural Networks

1 Introduction

Major Depression Disorder (MDD) is a psychiatric disorder defined as a state of low mood with a significantly higher level of duration/severity. It negatively impacts one's day to day life, causing people to become reluctant or unable to perform everyday activities, which can negatively affect a person's sleeping, sense of well-being, behaviour, feelings, etc. [1]. In extreme cases it can lead to suicide, which is the leading cause of death for men under 50 in the UK [2]. Depression is currently the most prevalent mental health disorder and the leading cause of disability in developed countries. A correct and early diagnosis can be vital to provide the right mental health support at the right time. It facilitates communication between (potential) patients and health professionals about the support and services they need [3] and is the key to choosing the correct intervention for treating patients.

Standard clinical depression assessment techniques can be subjective because these depend almost entirely on the health professional's own understanding of the individual's verbal psychological report, e.g. clinical interview and questionnaires completed by patients or caregivers [4]. In addition, this is often a lengthy procedure which hinders access to early treatment. In the UK it has been reported that more than half of the patients have to wait at least 3 months before receiving talking treatment [5]. Sometimes the relevant patient information or mental health experts may not be accessible, which results in many patients missing the best chance for preventing or treating their depression at early stages of depression. This is problematic, because correct early diagnosis is an important factor in the treatment of depression. To improve this, automatic objective assessment methods to aid monitoring and diagnosis have been widely explored in recent years.

There is convergent psychological evidence [6], [7], [8], [9], [10], [11], [12], [13] that depression is marked by non-verbal objective cues related to head movements, facial expressions and gaze [14], [15], which can be automatically detected and analyzed without the intervention of clinicians. Building an automatic system based on such cues would not only provide an objective and repeatable evaluation but would also help alleviate key problems around cost and time requirements [16].

Most current vision-based approaches to automatic depression analysis [17], [18], [19], [20], [21] base their prediction on the non-verbal facial behaviours of participants during an interview. There remain several challenges to achieve actionable results in this scenario, and our proposed approach mainly focus on addressing three of them. The first challenge is that the lengths of interview videos are usually variable, with the duration of the longest video sometimes several times longer than the shortest one. Yet most Machine Learning models require fixed-size input. The first research question we aim to answer is how to

- *Siyang Song and Linlin Shen are with Computer Vision Institute, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China. Linlin Shen is also with Shenzhen Institute of Artificial Intelligence and Robotics for Society, PR China and the National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China. Corresponding Author: Prof. Linlin Shen. E-mail: llshen@szu.edu.cn*
- *Siyang Song, Shashank Jaiswal and Michel Valstar are with the Computer Vision Lab, School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, United Kingdom. E-mail: {siyang.song, shashank.jaiswal1, michel.valstar} @nottingham.ac.uk*

Manuscript received July 29, 2019; revised January 11, 2020.

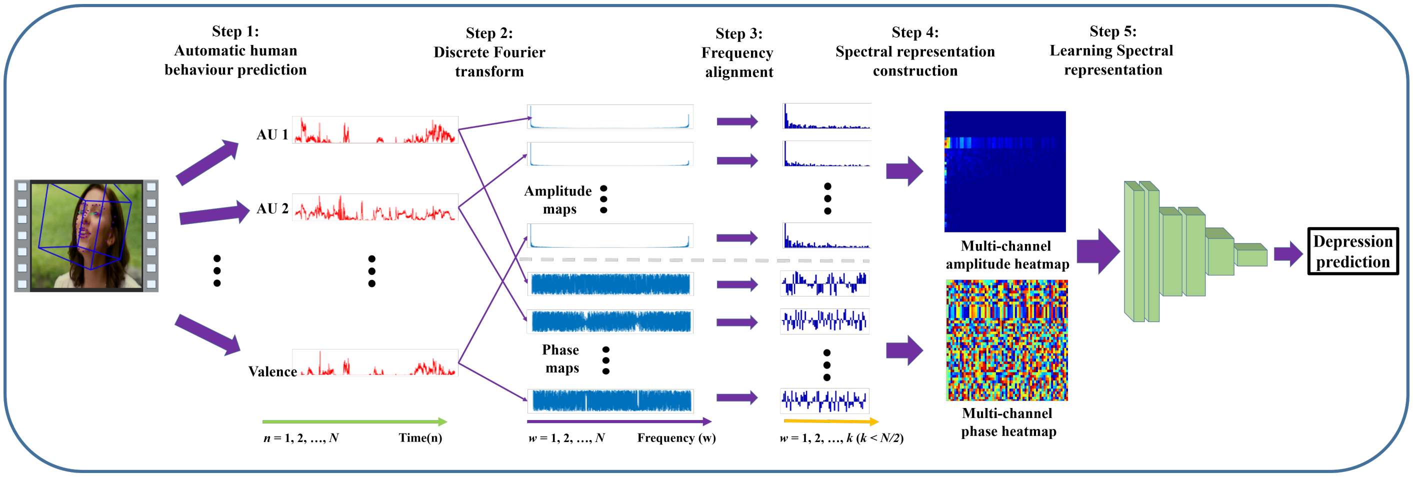


Fig. 1. The pipeline of our approach. Our approach starts with using low dimensional multi-channel human behaviour time-series data to represent videos (**Step 1, Sec. 3.1**), and then converts them to spectral signal consisting of multiple frequency information of all frames (**Step 2, Sec. 3.2.1**). Since spectral signals are symmetric, we only keep the first half of them. We then implement frequency alignment by removing high frequency components and obtaining common frequencies for human behaviours of all videos (**Step 3, Sec. 3.2.2**). Finally, we combine aligned spectral signals of all human behaviours (**Step 4, Sec. 3.2.3**) into and feed them to ML models for depression analysis (**Step 5, Sec. 3.3**).

encode information from variable length videos into a fixed-size video representation while retaining as much relevant information as possible. The second challenge is that while a number of studies have indicated that features such as facial expressions [6], [10], [22], head movements [23], [24], etc., are valuable for depression analysis, it is unknown how best to encode the temporal pattern of such features. Hence the second research question we aim to answer is: what is the most optimal way to extract such features preserving as much temporal information as possible. In particular, we are seeking feature descriptors that encode temporal patterns at multiple temporal scales, i.e. from short-term to long-term. The third challenge we address concerns the context under which people are observed. Depression interviews are usually made up of several tasks, e.g. reading paragraphs [25], [26], answering questions [27], etc. As a result, different tasks would trigger different responses from participants, leading to different facial behaviours. We define the third research question as how to learn such context specific behaviour and combine them for depression analysis.

Regarding the first challenge, one popular solution is to predict depression for each frame or short video segment, and then fuse the predictions using either a simple average [18], [25], [26], [28], linear regression [29] or Long-short-term-memory Network (LSTMs) [30]. However, except for memory-based neural networks such as LSTMs (which needs very large datasets), these approaches ignore long-term temporal behavior patterns of participants, which may better predict depression because behaviour extracted from a single frame or a short segment can be ambiguous and explained by various causes, e.g. a smile may be caused by feeling happy or feeling helpless. Also, the same short-term behaviours can be expressed by subjects with different levels of depression. In other words, depression levels can be more reliably described using the whole video rather than short segments of the video.

Alternatively, some studies constructed video-level descriptors by fusing frame/segment-level representations. To this end, one could consider re-sampling per-frame representations of a video (which can be a multi-channel time-series data) to a fixed length by using interpolation, Dynamic Time Warping (DTW), etc. However, this approach will distort the original signals. To avoid distortion, other studies employed fixed-size histogram or other statistics to summarize the distribution of representations. Specifically, they generate video-level descriptors by computing statistics of features [31], [32], [33], [34], using Gaussian Mixture Model (GMM) [35], [36], [37], [38], [39], [40] or fisher vector [38], [41], etc. Although these methods summarize undistorted information, temporal relations between segments/frames, such as the order of events, are lost after creating the statistics.

To deal with the second challenge, i.e. retaining multi-scale temporal dynamics, recent studies [20], [28], [42] usually divide each video into a series of short segments (ranging from 5 frames to a few seconds), and then extract temporal features from them. However, the optimum duration of the segments, which determines the temporal scale, is hard to determine. Such approaches only encodes a single-scale or possibly a small number of temporal scales, which ignores long-term temporal dynamics.

For the third challenge, while a few related studies [33], [43] are available, to the best of our knowledge, there is no study which systematically investigated what is the optimum way to use context specific behaviours for depression analysis. In this paper we make a systematic start at studying context by investigating the effects of a number of user tasks, as well as the effect of the duration of the most promising task.

In this paper, we aim to address these three challenges, avoiding drawbacks of previous works. Our approach consists of employing multiple, objective, visual and non-verbal human behaviour attributes that are easily interpreted by both people and machines, to wit Facial Action Units (AUs), head pose and gaze directions. We refer to these as *behaviour primitives*. By concatenating these frame-wise descriptors we obtain a multi-channel time-series describing the visual expressive human behaviour signal. To obtain a multi-scale, length-independent representation we propose two simple spectral representations that encode the human behaviour signal of the whole video. The proposed spectral representations contain video-level behaviour information in the frequency domain, where each frequency component stands for a unique scale of dynamics. We further employ two frequency alignment methods to create spectral representations of equal size and frequency coverage, regardless of variation in the length of input videos. Finally, we feed spectral representations to standard ML models (ANNs and CNNs), allowing dynamics of human behaviour obtained from multiple channels, to be jointly learned for prediction of depression severity. To investigate the third challenge, we conducted a series of experiments to compare the depression prediction results yielded by a series of tasks available in a benchmark dataset, as well as the results achieved by different fusion strategies, to wit, input-level fusion, feature-level fusion and decision level fusion. The overview of the proposed method is depicted in Fig. 1. In summary, the main novelty and contributions of this paper are listed as follows:

1) We propose a novel Fourier Transform-based approach that converts long and variable length time-series data to short and fixed-size spectral representations, which can be easily used with standard Machine Learning techniques.
2) The proposed spectral representations which encodes multi-scale video-level temporal dynamics of human behaviours, are shown to be useful for automatic depression analysis.
3) We investigate the influence of each automatically detected behaviour primitive on depression analysis, and found that AU4, AU12, AU15 and AU17 are useful for estimating depression severity, supporting existing evidence.
4) We investigate the influence of interview contents on depression analysis, and found that different interview tasks can result in completely different depression predictions.
5) We attain state of the art results for the estimation of depression severity when evaluating our proposed approach on the AVEC 2013 and AVEC 2014 datasets.

2 Related Work

2.1 The relationship between non-verbal cues and depression

In the past decade, many psychological studies have researched the relationship between non-verbal human behaviours and depression. Among these studies, a finding that depression is usually accompanied by reduced positive facial expressions, has been frequently concluded [10], [11], [12], [13], [22], [22], [44]. There is also some evidence that depression is associated with reduction in general facial expressiveness [12], [45] and head movements [23], [24]. Ellgring et al. [6] summarized typical symptoms of depression in terms of facial expressions, which indicates that depression is not only associated with sad facial expression but also with *"a total lack of facial expression corresponding to the lack of affective experience"*. Regarding the negative facial expressions, researchers have conflicting conclusions. While [46], [47], [48] argues that depression is marked by increased negative expressions, other studies [12], [45] found that depressed individuals are more likely to experience reduced negative expressions.

As a consequence, several studies have tried to apply such non-verbal cues to recognize depression. Cohn et al. [4] explored the feasibility of using audio and visual non-verbal cues for depression classification. They fed three different kinds of non-verbal behavioural features, i.e. manually annotated Facial Action Units (AUs), Active Appearance Model (AAM) features and vocal prosody features, to Support Vector Machines (SVM), individually. The results show that all of them were informative for detecting depression, with facial AUs achieving the best accuracy of 88%. The aforementioned findings suggest that automatic facial behaviour analysis could be useful for automatic depression analysis. Girard et al. [9] specifically investigated the relationship between depression and non-verbal facial behaviours, e.g., AUs and head poses, using manual and automatic systems. The results from both systems showed that participants with high depression severity presented fewer affiliative facial expressions (AUs 12 and 15), more non-affiliative facial expressions (AU 14) and diminished head motion.

2.2 Automatic depression analysis

Hand-crafted approaches In the past decade, automatic depression analysis has attracted a lot of attention, and a series of challenges have been organized [25], [26], [49], [50]. Early works [29], [31], [32], [51] generally use traditional Machine Learning models, e.g. Support Vector Machine Regression (SVR) [25], [33], decision tree [21], [43], [52], Logistic regression [53], etc., to predict depression from hand-crafted features (Local Binary Pattern (LBP) [38], [41], Low-Level Descriptor (LLD) [21], [34], [43], Histogram of oriented gradients (HOG) [26], etc). For example, Meng et al. [29] extracted LBP and EOH as visual features and LLD as audio features, and applied Motion History Histogram (MHH) to extract dynamics from short video segments. These features were fused together using Partial Least Square (PLS) regression to predict depression. The video-level decision is then made by combining the decisions from all segments using linear opinion pool. Gupta et al. [32] used LBP-TOP to summarize short-term temporal information and combined it with motion features and facial landmarks. A feature selection step is applied and the selected features are then used train a SVR model. [41] is another typical approach based on the combination of hand-crafted features and traditional ML models. This work extended the LBP-TOP feature to MRLBP-TOP for extracting short-term dynamics and then applied Fisher Vector to aggregate them.

Williamson et al. [35], [54] were the winners of the AVEC 2013 [25] and AVEC 2014 depression challenge [26]. Their methods were based on audio data and utilized formant frequencies and delta-mel-cepstra to represent underlying changes in vocal tract shape and dynamics. After that, by exploring the correlations between these features and using PCA, an 11-dimensional feature vector (five principal components for the formant domain and six principal components for the delta-mel-cepstral domain) is obtained. Finally, a Gaussian Staircase Model which is an extension of the Gaussian Mixture Model(GMM), was introduced and used as the regression model. Another approach proposed by Cummins et al. [36] is also based on GMM where a GMM-UBM model was employed to learn features that contain both audio and visual information. Jain et al. [37] also extracted LBP-TOP, HOG, HOF and MBH features and used GMM (Fisher Vector) to fuse the features from multiple video segments. Another GMM-based model was employed by Nasir et al. [55] where they proposed to use i-vector to learn several audio features such as TECC and MFCC.

Deep Learning approaches Due to the recent advances in deep learning, most current approaches build on Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Ma et al. [56] proposed DeepAudioNet for audio-based depression classification, which combines CNN with Long-Short-Term-Memory Networks (LSTMs). Most vision-based methods divide videos into several equal length segments, and extract deep learning features from each segment independently. Al Jazaery et al. [42] employed C3D network to extract short-term dynamic depression-related features from short video segments. Then, these features were fed to a Recurrent Neural Network (RNN) to make segment-level predictions. The final prediction was obtained by averaging predictions from all segments. Similar approach was proposed by Melo et al. [28] using 3D CNNs. To identify the salient facial region for depressed people, Zhou et al. [18] proposed DepressNet to learn depression representations with visual explanation. In this method, the facial region that is most informative to depression was highlighted and used to predict depression at frame level. The video level depression score was computed by averaging the scores from all frames. Recently, Haque et al. [57] employed Causal Convolutional Networks [58] to learn from audio, text and 3D facial landmarks to predict depression severity.

Besides learning depression directly from images, some approaches attempted learning depression severity from higher level video representations. Yang et al. [21] proposed selecting several equal length segments in each video to balance the number of depressed and non-depressed training examples. They also proposed a Histogram of Displacement Range (HDR) method that records the dynamics of facial landmarks in a video segment. They used CNNs to learn deep features from hand-crafted audio and video descriptors and the final decision is made by fusing the predictions from audio, video and text features using a decision tree. To predict depression directly from variable length videos, the previous version of our work [59], [60] used several human behaviour primitives to represent each frame, reducing a video to a multi-channel time-series data. In this paper, besides applying Fourier Transform to convert multi-channel time-series data to the frequency domain, we further explain how to align the frequencies of converted spectral signals to use a fixed set of frequencies to represent any video. In addition, we also investigate the influence of behaviours and task contents on depression analysis in this paper.

3 The proposed approach

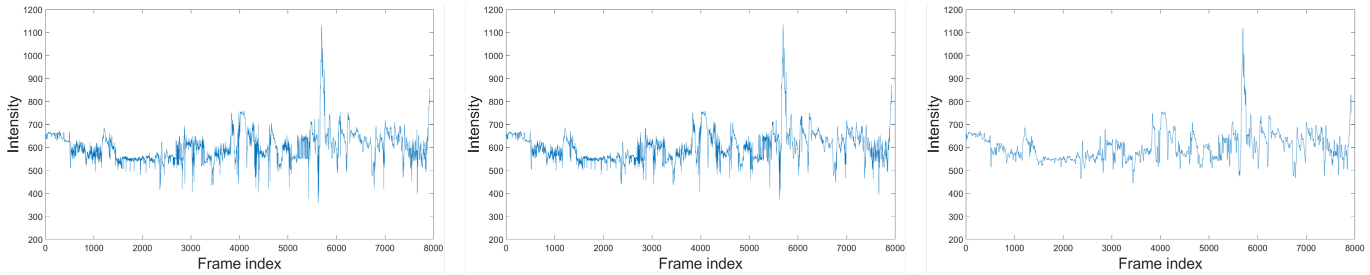
In this section, we describe a novel video-based automatic depression analysis approach that can extract fixed-size descriptors from variable length videos, encoding multi-scale temporal information. To achieve this, we first extract a set of automatically detected human behaviour primitives to represent a video, allowing the high dimensional videos to be significantly reduced to a low dimensional multi-channel time-series signal (Sec. 3.1). In Sec. 3.2, we propose two spectral representations as the video-level descriptors for multi-channel behaviour signals, which can not only encode a time-series data of arbitrary length into fixed-size representations but also retain multi-scale temporal information from the original time-series data. Finally, we show how to apply the generated spectral representations to depression analysis (Sec. 3.3).

Compared to other recent methods, the main advantages of our approach are: 1) it can convert long and variable length time-series data to a short and fixed-size representation, allowing information from the whole video to be used for analysis. It differs from [21], [52] in the sense that they only use some segments from videos for analysis. Our representations contains multi-scale video-level temporal information, in contrast to [33], [34], [39], [41] where only a fixed-length time window is used to encode single-scale short-term dynamics, thereby losing temporal information at other scales.

3.1 Human behaviour primitives extraction

In aiming to construct a video-level descriptor, the first task is to reduce the dimensionality. Current studies either extract hand-crafted features [38], [41], [61] or deep-learned features [18], [20], [42] to represent each frame or short video segment. Traditional hand-crafted features, e.g. HOG, LBP, etc, are not specifically designed for facial behaviour applications, consequently, they are not the most optimum representation for depression application. On the other hand, as summarized in Sec. 2, previous psychological and computer vision studies suggested that depression is marked by non-verbal visual cues. Motivated by this, we propose to use facial behaviour attributes, including AUs, gaze direction and head pose as frame-wise descriptors. In particular, we use OpenFace 2.0 [62] to automatically detect intensities of 17 different AUs, gaze directions and head pose, resulting in a 29-channel human behaviour time-series data for each video (17 corresponding to AUs, 6 corresponding to gaze direction from each eye and 6 corresponding to head-pose).

Compared to previously used hand-crafted and deep-learned features, these human behaviour descriptors have several advantages. Firstly, they are more interpretable, as they have a clearly understood meaning and are low-dimensional; Secondly, their extraction is modular, because standard facial attribute detection software, frequently trained on very large databases, can be used for different people in different scenarios; Thirdly, they are objective, as their values are independent of the subjects' identities, preventing the final predictions from being affected by bias related to gender, age, race, etc.; Fourthly, the proposed behaviour descriptors have much lower dimensionality (31-



Original signal **Remove 3000 highest frequencies** **Remove 7000 highest frequencies**

Fig. 2. Reconstructed time-series signals after removing high frequency components. The original signal has 7923 frames and its spectral signal also has 7923 frequencies.

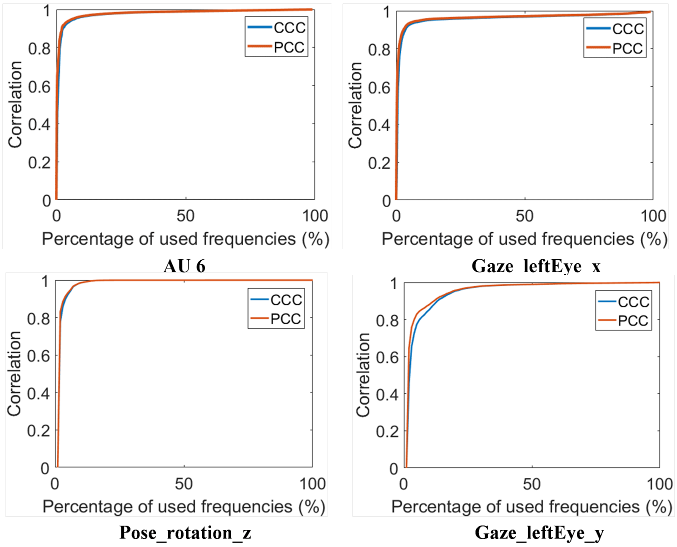


Fig. 3. Example of the average correlation between reconstructed behaviour signals and original behaviour signals as a function of the percentage of used frequencies. It can be observed that even after removing more than 90% of high frequency components, the reconstructed signals still have significant correlation with the original signals. (The Pose_rotation_z is the least useful behaviour primitive for depression analysis; the Gaze_leftEye_y achieved lowest correlation (CCC) with the original signals when removing 90% high frequency components)

D) than the traditional hand-crafted features and deep-learned representations.

3.2 Spectral representation for human behaviour primitives

To construct a spectral representation for multi-channel time-series data, we first transform each time-series to the frequency domain. We further propose two frequency alignment methods so that the spectral representation of each video (potentially of different lengths) represents the same frequencies. Finally, we also propose two ways to combine spectral representations of all behaviour primitives to produce a single representation for a given video. In this paper, we define the $f_c^m(n)$ as the c_{th} behaviour time-series signal in m_{th} video.

3.2.1 *Encoding multi-scale video-level dynamics*

Given that depression causes changes in behaviour which can be represented by time-series signals, temporal patterns are significant. Depression causes long-term changes in behaviour and so we aim to extract video-level features which can encode temporal patterns including long-term temporal information.

We use the Fourier Transform (FT) to convert time-series signals representing each behaviour primitive to the frequency domain. The resulting spectral representation is a decomposition of the original time-series into its constituent frequencies. Let $f(x)$ be a time-series signal corresponding to a behaviour primitive, then the Fourier Transform can convert it to a spectral representation $F(w)$

$$F(w) = \int_{-\infty}^{\infty} f(x) e^{-(2\pi i x w)/N} dx \tag{1}$$

where w can be any real number and $F(w)$ is a complex function that can be re-written as

$$\begin{aligned} F(w) &= \int_{-\infty}^{\infty} f(x)(\cos((2\pi i x w)/N) - i\sin((2\pi i x w)/N))dx \\ &= \int_{-\infty}^{\infty} (\mathrm{Re}(fc(x)) + i\mathrm{Im}(fs(x))) \\ &= \mathrm{Re}(F(w)) + i\mathrm{Im}(F(w)) \end{aligned} \tag{2}$$

where $fc(x)$ and $fs(x)$ denote $f(x)\cos((2\pi i x w)/N)$ and $-f(x)\sin((2\pi i x w)/N)$, respectively. $R(F(w))$ is the real part of $F(w)$ and $Im(F(w))$ is the corresponding imaginary part of $F(w)$. Here, w determines the frequency $(2\pi w)/N$ that $F(w)$ represents. Consequently, the spectral representation $F(w), w \in [-\infty, \infty]$ contains information from all frequencies present in $f(x)$.

In our application, each video is made up of a series of frames, resulting in one discrete time-series signal for each behaviour primitive. We therefore apply Discrete Fourier Transform (DFT) to the behaviour signal $f_c(n)$, where $c = 1, 2, \cdots, C$ denotes the behaviour primitive's index and $n = 1, 2, \cdots, N$ denotes the frame index, as given below:

$$\begin{aligned} F_c(w) &= \sum_{n=0}^{N-1} f_c(n) e^{-\frac{2\pi i}{N} wn} \\ &= \sum_{n=0}^{N-1} f_c(n)[\cos(2\pi wn/N) - i\sin(2\pi wn/N)] \\ &= \sum_{n=0}^{N-1} (\mathrm{Re}(fc_c(n)) + i\mathrm{Im}(fs_c(n))) \\ &= \mathrm{Re}(F_c(w)) + i\mathrm{Im}(F_c(w)) \end{aligned} \tag{3}$$

where $f_c(n)$ is the time-series signal of c_{th} behaviour, which consists of N frames and $F_c(w)$ is the DFT of the signal $f_c(n)$ at frequency w, where $w = 0, 1, 2 \cdots, W-1$.

As we can see from Eq. 3, each frequency component is computed from all frames of the $f_c(n)$. This is to say, each component in the spectral signal summarizes a single frequency information present in the whole video. Therefore, the spectral signal contains information corresponding to W frequencies given by $2\pi w/N, w = 0, 1, 2, \cdots W-1$. These components encode different types of behaviour dynamics, i.e. high frequency components represent sharp behavioural changes and low frequency components represent more gradual changes in behaviour. As a result, the produced spectral signal can be said to summarize multi-scale temporal information of the whole video. Here, we set the number of discrete frequency components W in $F_c(w)$ to be the same as N in order to completely summarize the information contained in the discrete time-series data $f_c(n)$ (It is well known that the $f_c(n)$ can be fully reconstructed from $F_c(w)$ if $W = N$).

3.2.2 Frequency alignment

As mentioned above, a time-series behaviour signal of N frames can be converted to a spectral signal that has $W = N$ frequency components. Thus, the spectral signals of variable-length videos will have different number of components, which would again lead to feature representations of varying dimensionality. To make them equal, we first note that the spectral signals of time-series data are always symmetric around their central frequency $W/2$, i.e. if $F(w) = Re(w) + iIm(w)$ and $F(W-w) = Re(W-w) + iIm(W-w)$, then $Re(w) = Re(W-w), Im(w) = -Im(W-w)$. This means that the first $W/2$ components of the spectral signal can fully represent the information contained in $f(n)$. Also, as facial actions are continuous and smooth processes, high-frequency information usually represents noise or outliers caused by e.g. incorrectly detected faces, errors in facial points localization or AU intensity estimation etc. In practice, after removing the high-frequency information, the reduced spectral signal can still represent the original time-series data well, as applying the inverse DFT to spectral signals can recover most of the information present in the original time-series data. This is illustrated in Fig. 2 and Fig. 3. In both figures, we replace all unused frequency components by zeros.

Motivated by this, our approach only keeps the first $W/2$ components of spectral signals. Then, components corresponding to high frequencies are also removed. Since our goal is to generate video-level spectral representation of the same size for variable length time-series data, one may consider to keep the first K lowest frequencies of spectral signals for all videos, with $K < W/2$. However, the w_{th} component in videos of different lengths will represent different frequencies. Consider two time-series signals $f^1(n)$ and $f^2(n)$ of length N_1 and N_2 respectively. Also consider their corresponding spectral representations as $F^1(w)$ and $F^2(w)$ respectively. If $N_1 \neq N_2$, the w_{th} component $(0 < w < N_1/2, N_2/2)$ of the spectral signal $F^1(w)$ denotes the DFT value at frequency $2\pi w/N_1$ while the w_{th} component of $F^2(w)$ denotes the DFT value at frequency $2\pi w/N_2$. Clearly, $2\pi w/N_1 \neq 2\pi w/N_2$, and thus the w_{th} component of spectral signals $F^1(w)$ and $F^2(w)$ do not represent the same frequency. In order to resolve the above problem of misaligned frequencies, we propose the following two solutions:

Solution 1: Zero-padding is a common method often used to increase the frequency resolution after Fourier Transformation of a discreet time series. In this method, zeros are appended to the time-series data to increase its length, allowing the DFT of this time-series data to have more frequency components. In particular, the frequency resolution W of the spectral signal is equal to the number of frames N in the original time-series data. By padding with zeros, we add N_{add} zeros at the end of the original time-series to create a new time-series of length $N + N_{add}$. Consequently, the spectral signal of the new time-series will have $W + N_{add}$ frequency components. Please see [63] for detailed theoretical explanation of this method. In this paper, we first obtain the total number of frames in the longest video of the dataset. Then, we add zeros to the behaviour signals extracted from the rest of the videos, making all behaviour signals to have the same length as the longest video. Consequently, the spectral signals of all zero-padded time-series behaviour signals will have the same resolution. By further selecting only the first K components of each spectral signal, the dimensionality can be significantly reduced.

Solution 2: Although zero-padding can increase the frequency resolution of spectral signals, the values of the increased frequency components are estimated. Moreover, the multi-channel facial behaviour time-series signals added by zero-padding are zero-signal. This strategy assumes that the facial status in the added frames is neutral and remains unchanged, which is not correct. Therefore, the extended multi-channel time-series signal cannot accurately represent the facial behaviour patterns of the corresponding person and the values of the increased frequency components are estimations only. To avoid this, our second solution extracts fixed-size spectral signals from variable length time-series data by choosing k common frequencies from the spectral signals obtained from each video. In this case, the values of k chosen frequencies are obtained from the original signal rather than an extended signal. Hence, each component in the produced representation represents the accurate value rather than estimated value of the corresponding frequency. It should be noted that the advantage of this method is at the cost that the spectral signals gets downsampled thereby losing some information. Assuming that there are M time-series signals $f^1, f^2, \cdots f^M$ corresponding to M variable length videos, the proposed solution follows following steps:

1) Choose a fixed frequency resolution R, i.e. the number of frequency components used to represent each time-series data, and then shorten the time-series, reducing the total number of frames in the original time-series signal $f^m(n)$ from N_m to $N_m - (N_m \bmod R)$ frames, which is a multiple of R, resulting in a slightly shorter time-series signal $S(f^m(n))$. In practice, we remove the first $(N_m \bmod R)/2$ frames and the last $(N_m \bmod R)/2$ frames from each video. In our experiments, N_m

was chosen as 100 (task-based experiments) or 500 (experiments on whole videos of AVEC 2013), which means the maximum length of removed video contents were less than 4 seconds and 17 seconds, respectively, (average 1.2 seconds and 6.6 seconds in our experiments, respectively, while the average full lengths of the videos are about 189 seconds and 961 seconds).

2) Each time-series $S(f^m(n))$ is converted to its spectral signal $S(F^m(w))$ using Equation (3). Since the number of frequency components is equal to the number of frames, the number of frequency components in $S(F^m(w))$ will also be multiple of R, which can be defined as $W_m = (t_m \times R), m = 1, 2, \cdots, M$. As a result, the frequencies represented in each spectral signal can be denoted as $2\pi w_m/(t_m \times R), w_m = 0, 1, 2, \cdots, t_m \times (R-1)$.

3) As the number of frequencies in each spectral signal is a multiple of R, all of them would contain the same R components whose frequencies are given by:

$$\begin{aligned} n_f(m) &= 2\pi w_m(r)/W_m \\ &= 2\pi r \times t_m/(R \times t_m) \\ &= 2\pi r/R \end{aligned} \tag{4}$$

where $r = 0, 1, 2, \cdots (R-1)$. It is clear that the R selected frequencies are independent of t_m, and these R frequencies, i.e. $2\pi \times 0/R, 2\pi \times 1/R, 2\pi \times 2/R, \cdots, 2\pi \times (R-1)/R$, are encoded in all spectral signals. This process is also illustrated in Fig. 4. Finally, we remove those high frequency components and only keep the first K components.

As a result, solution 2 can not only align the frequencies of variable length time-series signals, but also prevents distortion of the aligned spectral signals.

3.2.3 Spectrum representations

After obtaining aligned spectral signals corresponding to each behaviour primitive, we describe two different methods to construct a fixed-size joint representation so that all behaviour spectral signals can easily be used as input features for standard ML techniques.

Assuming that C behaviour primitives are extracted from each frame, we produce C aligned spectral signals consisting of K frequencies for each video. Since the values in spectral signals are complex numbers, we convert each of them to two spectrum maps in the real domain: an amplitude map and a phase map, where the amplitude map can be computed by

$$|F_c^m(w)|/N = \sqrt{\mathrm{Re}_c^m(w)^2 + \mathrm{Im}_c^m(w)^2}/N \tag{5}$$

and the phase map can be computed by

$$\arg(F_c^m(w)) = \arctan\frac{\mathrm{Im}_c^m(w)}{\mathrm{Re}_c^m(w)} \tag{6}$$

where $\mathrm{Re}_c^m(w)$ and $\mathrm{Im}_c^m(w)$ are the real and imaginary part of $F_c^m(w)$ respectively, as defined in Equation(3). Hence, C amplitude maps and C phase maps are extracted from a video, all of which have K frequencies. We further propose the following two methods to combine them:

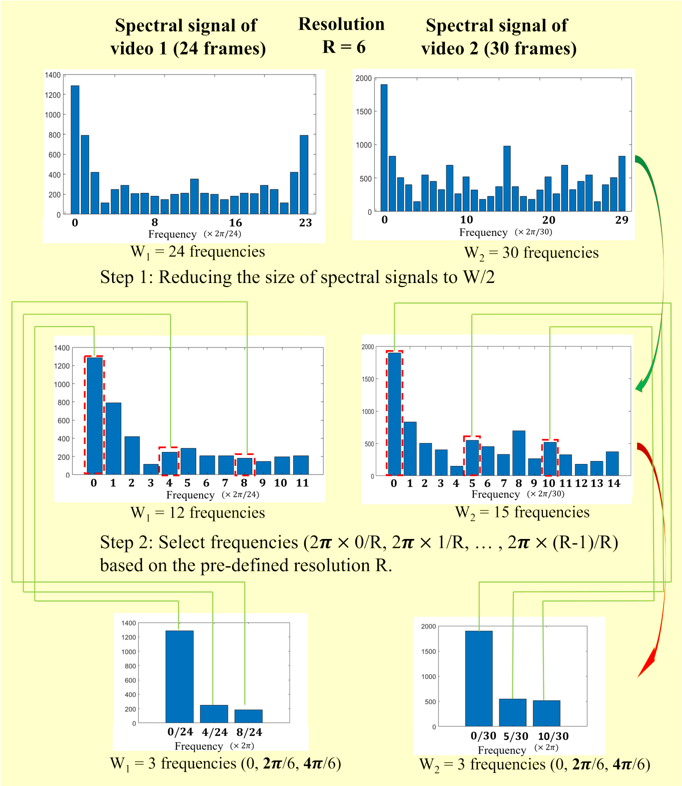


Fig. 4. Frequency selection (Step 3 of solution 2): After the DFT, the second half of the spectral signals are removed as they are symmetric.

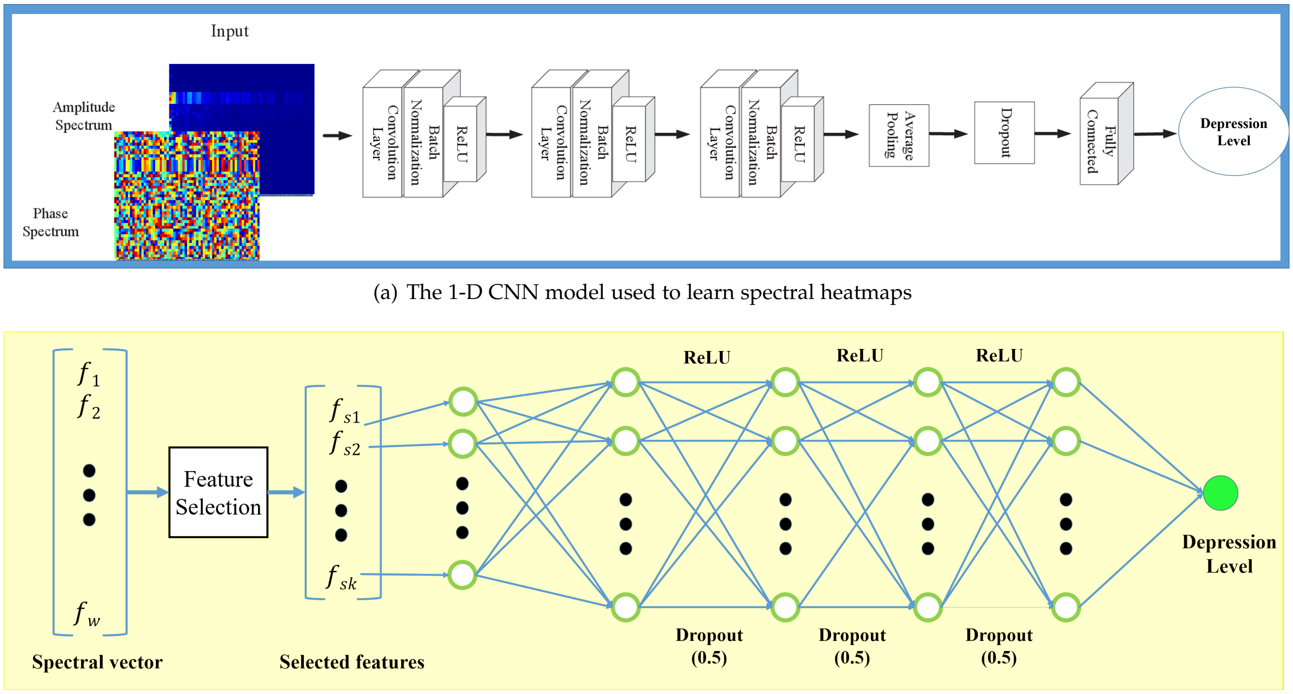
1) **Spectral heatmap**. A $C \times K$ multi-channel amplitude spectrum map and a $C \times K$ multi-channel phase spectrum map. In both maps, each row represents an amplitude map or a phase map of a single behaviour spectral signal while each column represents a frequency. In this paper, we combine two spectrum maps as a two-channel spectral heatmap.

2) **Spectral vector**. A 1-D vector that concatenates $C \times K$ amplitude features and $C \times K$ phase features from all behaviour primitives. As a consequence, the concatenated vector contains $C \times K \times 2$ components.

It is clear that both representations encode information from all human behaviour signals. Also, their fixed size makes them suitable for use with standard ML techniques.

3.3 Learning spectral representations

Inspired by recent advances in deep learning for multi-channel signal processing including audio feature processing [56], we employed a 1-D CNN structure [64] that has been frequently used in the multi-channel time-series data analysis, to extract features from **spectral heatmaps**. The reason behind this is that the rows in the heatmaps, which represent a set of behaviour primitives, have no natural ordering, spatial or otherwise. Therefore, standard 2-D CNNs may not be suitable. Hence, the proposed spectral heatmaps are treated as multi-channel 1-D data and 1-D CNNs are used to learn depression prediction networks. As shown



(a) The 1-D CNN model used to learn spectral heatmaps

(b) The ANN model used to learn spectral vectors

Fig. 5. The ML models employed in this paper.

in Fig. 5(a), our CNN architecture is made up of three Conv-Batch-ReLU blocks, where each block contains a 1-D convolution layer followed by a batch normalization layer and a ReLU layer. In particular, each convolution layers consists of 128 filters of kernel size 7×1, 128 filters of kernel size 5×1, and 64 filters of kernel size 3×1, respectively. After that, a channel-level average pooling layer is employed to obtain a 1-D feature from each feature map, producing a 64-D deep feature. Finally, a fully connected layer with 64 input neurons, a dropout layer [65] (probability factor $p = 0.5$) and an output layer of one neuron are used at the top of the average pooling layer to predict depression levels.

For **spectral vectors**, we propose to use an Artificial Neural Network (ANN) structure used in [60], which consists of four fully-connected hidden layers displayed in Fig.5(b) for regression. The dimension of spectral vectors is usually much higher compared to the amount of training data (usually less than 200 training and validation examples [25], [26], [27]). This may lead to model overfitting. In order to avoid this, we introduce Correlation-based Feature Selection (CFS) [66] to reduce the dimensionality. CFS is a popular feature selection technique which only selects those features which are highly correlated with the output variable but uncorrelated with each other, thereby giving a very compact set of useful features. In our implementation of CFS, we employed Pearson's linear correlation coefficient to measure the correlations. Considering that the distribution of training labels are usually unbalanced, we group them into b classes based on their depression severity labels and apply a voted version of CFS to decide the final feature set. The procedure of V-CFS is explained in Algorithm (1).

Algorithm 1 Procedure of V-CFS

```
1: Divide the training set into $n$ subsets based on their labels, where each subset may contain a different number of examples.
2: Select the same number ($k$) of examples from each of the subset, resulting in $k \times n$ selected examples;
3: Apply CFS to the selected examples
4: Repeat Step 2 and Step 3 $t$ times, resulting in $s$ selected features;
5: Vote on all features and rank them in descending order based on frequency;
6: Select top ranked features as the final feature set.
```

4 EXPERIMENTS

In this section, we first describe the experimental settings, including the datasets (Sec.4.1), pre-processing (Sec.4.2), model training details (Sec.4.3) and the performance metrics (Sec.4.4). Then, we describe the interactive behaviour studies which investigates the influence of behaviour primitives (Sec.4.5.1), task contents (Sec.4.5.2) as well as the length of videos (Sec.4.5.3) on depression analysis. The frequency selection was used to align frequencies for all interactive behaviour studies. Since our approach has two frequency alignment methods and two spectral representations, we also present ablation studies in Sec.4.6 to evaluate their performance. Finally, we also compare our best system to the state-of-the-art approaches (Sec.4.7).

4.1 Datasets

We conducted our studies on the AVEC 2013 [25] and AVEC 2014 [26] audio-visual depression corpus. In the AVEC 2016 dataset, the set of tasks completed by each participant

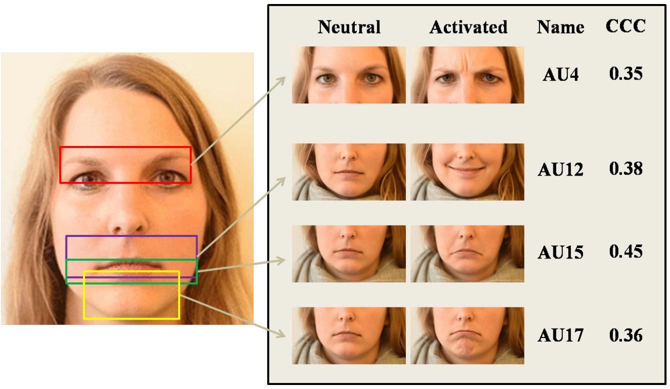


Fig. 6. Visualization of the most important facial actions and their face region for depression analysis.

are sometimes different. Because the analysis of behaviour is context-dependent and in order to avoid the negative impact of the variation of tasks completed by the participants, we decided against using the AVEC 2016 dataset for our experiments. The corpus used by the AVEC 2013 challenge contains 150 audio-visual clips. Each clip records a participant doing a series of tasks, including sustained vowel phonation, sustained loud vowel phonation, sustained smiling vowel phonation, speaking out loud while solving a task, counting from 1 to 10, etc. All participants are German speakers and each of them do the same tasks in the same order during the video recording. The length of these videos ranges from 20 minutes to 50 minutes with an average of 25 minutes. The audio-visual depression corpus used by AVEC 2014 challenge also contains 150 audio-visual clips. In contrast to AVEC 2013, AVEC 2014 contains two audio-visual files for each participant corresponding to two different tasks, i.e. Northwind and Freeform, resulting in much shorter length of each video. For both datasets, the frame rate of videos were set to 30 frames per second with resolution of 640 × 480, and each clip is labeled with a Beck-Depression Inventory (BDI II) score indicating the depression severity. These scores range from a minimum of 0 to a maximum of 63.

4.2 Pre-processing

In this paper, we employed the OpenFace 2.0 toolkit [62] to automatically detect intensities of 17 AUs (AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26 and AU45), 6 gaze direction descriptors and 6 head pose descriptors (detailed explanation is in 7) resulting in a 29-D frame-wise human behaviour representation. For frames in which no face was detected or if the confidence value of the detected face was small, no features were extracted and such frames were removed from analysis. To minimize the effects of participants identity, the values of all human behaviour primitives were normalized by subtracting their corresponding median values computed over the whole video.

4.3 Training details

To learn from spectral heatmaps, our CNN structure was fixed (illustrated in Sec. 3.3 and Fig. 5(a)) for all experiments described in this paper. We also employ ANNs consisting of four fully-connected hidden layers (illustrated in Sec. 3.3 and Fig. 5(b)) to learn from spectral vectors, where the size (ranged from 20 to 40) of each hidden layer was optimized for each experiment individually. For all networks, we used Adam [67] as the optimizer and MSE as the loss function. All training hyper-parameters for ANNs and CNNs, e.g. learning rate, beta 1, beta 2 etc, and were optimized on a validation set for each experiment individually. Other hyper-parameters of the network, e.g. the number of layers and the pooling method, were chosen based on the average validation results of multiple experiments. They were kept the same for all experiments.

The spectral feature extraction, feature selection and ANN training were implemented in MATLAB 2019, while the CNNs were implemented in Pytorch.

4.4 Performance metrics

To compare the performance of our approach to previous solutions, we adopt two metrics used by the previous AVEC challenges, i.e. root mean square error (RMSE) and mean absolute error (MAE), which are defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{7}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{8}$$

where f_i is the predicted depression severity and y_i is the corresponding ground-truth. Additionally, we also report correlations between predictions and ground-truth, based on Pearson Correlation Coefficient (PCC, Eq. 9) and Concordance Correlation Coefficient (CCC, Eq. 10). PCC can be defined as:

$$\text{PCC} = \frac{\text{cov}\,(f, y)}{\sigma_f \sigma_y} \tag{9}$$

where the cov is the covariance and σ_f, σ_y are the standard deviations. On the other hand, the Concordance Correlation Coefficient (CCC) is given by:

$$\text{CCC} = \frac{2\rho_{f,y}\sigma_f\sigma_y}{\sigma_f^2 + \sigma_y^2 + (\mu_f - \mu_y)^2}, \tag{10}$$

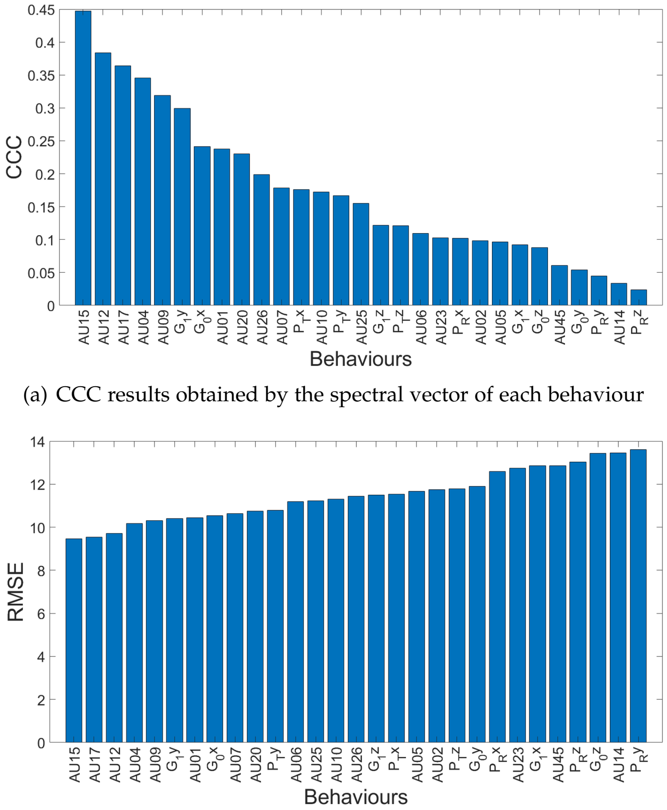
where μ_f and μ_y are the mean values of predictions and labels respectively while σ_x and σ_y are the corresponding standard deviations.

4.5 Interactive Behaviour Studies

4.5.1 *Analysis of facial behaviour primitives*

In this section, we independently evaluate the performance of each human behaviour primitive on depression severity estimation. To do so, we trained separate models from the spectral vectors of each behaviour primitive. The results are reported on the AVEC 2013 dataset.

As shown in Fig.7, individually using spectral vector of AU15, AU17, AU12, AU04 and AU09 intensities achieved decent performance, where CCC results are over 0.3 and RMSE results are less than 10. Particularly, AU15 yielded the best result among 29 behaviour primitives, with more



(a) CCC results obtained by the spectral vector of each behaviour

(b) RMSE results obtained by the spectral vector of each behaviour

Fig. 7. Depression severity estimation results obtained by spectral vector of each facial behaviour primitive, on AVEC 2013 dataset.(G_0x, G_0y, G_0z represent the gaze direction for left eye; G_0x, G_0y, G_0z represent the gaze direction for right eye; P_Tx, P_Ty, P_Tz represent the location of the head; P_Rx, P_Ry, P_Rz represent the rotation of the head. Please see [62] for details).

than 0.4 CCC. Head pose and gaze directions seemed less informative to depression, at least on their own, as five features corresponding to these ranked at the bottom end of both CCC and RMSE. We also analyzed the temporal activation patterns of the four most informative AUs (AU4, AU12, AU15 and AU17) through a number of statistical measures. We did this to gain some insight in the human interpretable differences of facial behaviour of severely depressed and non-depressed people. For this analysis, we re-framed the task as a binary classification problem between participants that have BDI score between $29 - 63$, which is defined as severely depressed according to BDI II questionnaire, and participants that have BDI score between $0 - 13$, which is defined as minimally-depressed. The findings are reported in Table 1, showing that people with depression tend to frequently display AU4 activation. The average duration and intensity of AU4 activation also tend to be higher in depressed people. On the other hand, the activation of AU12 was found to be less frequent in depressed people. In addition, they are also more likely to have shorter AU15 activation and longer AU17 activation. These results show that there is a significant amount of information present in some of these behaviour primitives which could be exploited for automatic depression analysis. Fig. 6 visualizes the most important facial regions for video-based depression analysis based on the aforementioned results.

Behav	Description	Correlation	Mean PCC
AU 4	Frequency of activation	Positive	0.36
AU 4	Mean activation intensity	Positive	0.30
AU 4	Median activation intensity	Positive	0.26
AU 17	Mean distance between two adjacent activation	Positive	0.26
AU 17	Median activation duration	Positive	0.25
AU 15	Median duration of activation	Negative	-0.25
AU 17	Frequency of activation	Negative	-0.24
AU 12	Frequency of activation	Negative	-0.23
AU 4	Mean distance between two adjacent activation	Negative	-0.20
AU 4	Standard deviation of activation durations	Positive	0.19
AU 17	Mean activation duration	Positive	0.18
AU 4	Median activation duration	Negative	-0.15

TABLE 1
Analysis of human interpretable temporal AU activation patterns

Table. 2 reports the results achieved by each single modality, e.g. AUs, gaze and head pose as well as their combinations, showing that out of the three visual cues, AUs achieve the best performance when used independently, and the best result were obtained by fusing all cues. To determine the added value of individual features, we conducted an experiment where the results shown in Fig. 7 were used to evaluate the system with increasing numbers of features added in a greedy approach, i.e. starting with the feature that has the highest predictive value when used on it's own and then the top-2 features, top-3 etc. We report on the results using a figure Fig. 8 to explain how performance improved with features from more behaviour primitives. While the performance has been fluctuated a bit at some points, it is still clear that they were increasing when more features were used.

Please note that these results only indicate the relationship between automatically detected behaviours and depression, which may be slightly different to the result achieved by using human annotated behaviour information. This is because the tools we used for behaviour detection are not 100% accurate and the errors in detection may affect the depression analysis results.

4.5.2 Task-based depression analysis

The subjects in the videos of AVEC-2013 and AVEC-2014 dataset were recorded while doing a set of predefined tasks. To investigate the influence of different tasks on the performance of depression models, we divided the videos of AVEC 2013 dataset into several segments based on the task topics, which are: Task 1. Sustained vowel pronunciation; Task 2. Problem solving while speaking out-loud; Task 3. Counting from 1 to 40; Task 4. Reading a text out-loud;

Behav	MAE	RMSE	PCC	CCC
AUs	6.92	8.22	0.67	0.56
HP	8.54	10.38	0.37	0.30
Gaze	7.51	9.04	0.49	0.35
AUs+HP	6.96	8.34	0.67	0.55
AUs+Gaze	6.85	8.29	0.74	0.66
HP+Gaze	7.49	9.15	0.52	0.37
AUs+HP+Gaze	6.68	8.10	0.75	0.68

TABLE 2
Analysis of the influence of individual behaviours, where HP denotes head poses.

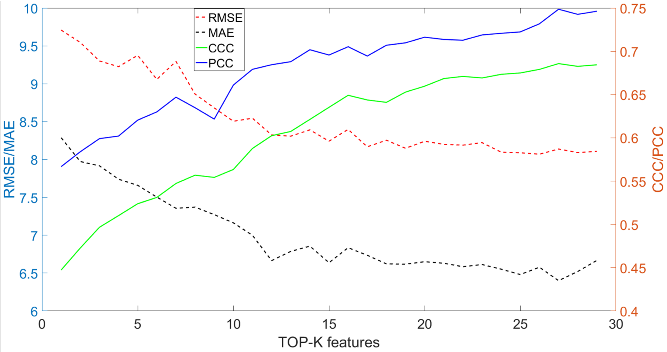


Fig. 8. The depression analysis results (AVEC 2013 dataset) for TOP-K behaviour primitives displayed in Fig.7(a).

Task 5. Singing (Task 5 was not completed by any of the participants); Task 6. Telling a story from one's childhood; Task 7. Telling a story based on a picture applying the Thematic Apperception Test (TAT). The order of these tasks in each video is constant, and is given by: 1. Task 4, 2. Task 3, 3. Task 7, 4. Task 6, 5. Task 1, 6. Task 2. To help other researchers in conducting similar studies, we have made the time-stamps and detailed description of these tasks, publicly available at [1]. In case of AVEC 2014 dataset there are already separate videos available for the two sub-tasks used in this dataset.

We conducted three types of experiments on both datasets: 1. single task-based experiments, where the model and the performance results were generated by using the video segment of each task separately; 2. Feature-level fusion of all tasks, where the video-level feature vector was obtained by concatenating features from the video segments of all tasks, and selected by V-CFS; 3. Decision-level fusion of all tasks, where the final predictions were obtained by combining the predictions from all tasks using linear regression. In addition, we also report the performance achieved using whole videos of AVEC 2013 dataset without considering the task boundaries. It should be noticed that only 35 training videos, 32 validation videos and 39 test videos contain all tasks. The results in this subsection are reported only on this subset of videos containing all tasks.

The results of all experiments achieved by are shown in Fig.9. It can be observed that the tasks contents have significant impact on the performance of our approach, as the results achieved by different tasks varied a lot; This can be explained by the fact that different tasks can trigger different facial behaviours, some of which are more informative than others, for detecting depression. Secondly, it is clear that the feature-level fusion and decision-level fusion of all tasks provided better results than using features from one task only, indicating that depression can be better analyzed by fusing information from multiple tasks. Thirdly, when comparing the three fusion strategies, i.e. input-level fusion (extract features from whole videos), feature-level fusion and decision-level fusion, the decision-level fusion achieved the best results and feature-level fusion outperformed the input-level fusion. This result suggests that modelling depression based on task segments can better predict depression severity than using whole videos without considering the time boundaries of tasks.

Alignment	MAE	RMSE	PCC	CCC
No alignment	9.09	11.16	0.29	0.11
Padding	7.66	9.59	0.53	0.33
Selection	7.44	9.46	0.52	0.39

TABLE 3
Comparison of average results generated by two frequency alignment on AVEC 2013 test set

Alignment	MAE	RMSE	PCC	CCC
No alignment	8.67	10.82	0.27	0.14
Padding	7.55	9.40	0.52	0.36
Selection	7.18	9.27	0.56	0.42

TABLE 4
Comparison of average results generated by two frequency alignment on AVEC 2014 test set

4.5.3 *Influence of video length*

In this section, we used between 10% and 100% of the video segments corresponding to task 4, in increments of 10%, to investigate the influence of video length on depression prediction. Since the task 4 asked participants to read the same text, the contents of this task are constant for all participants, which makes the analysis invariant to other factors, such as different story topics in Task 6, etc.

As we can see from Fig. 10, when the duration of videos is very short, the performance is low, However, when the video length is increased, longer-term behaviour becomes available for analysis. As a result, the depression estimation performance increases significantly, with the best result achieved by using the first 80% of task 4 video segments (the average length of the first 80% of videos is 338.7 seconds) in AVEC 2013 dataset.

4.6 Ablation Studies

4.6.1 *Comparison of frequency alignment methods*

As described in Section 3.2.2, we used two frequency alignment methods, i.e. zero-padding and frequency selection. In this section, we evaluate both methods on all the task-based and fusion experiments described in Section 4.5.2, and report the average performance achieved by each of

[1] https://github.com/SSYSteve/Human-behaviour-based-depression-analysis-using-hand-crafted-statistics-and-deep-learned

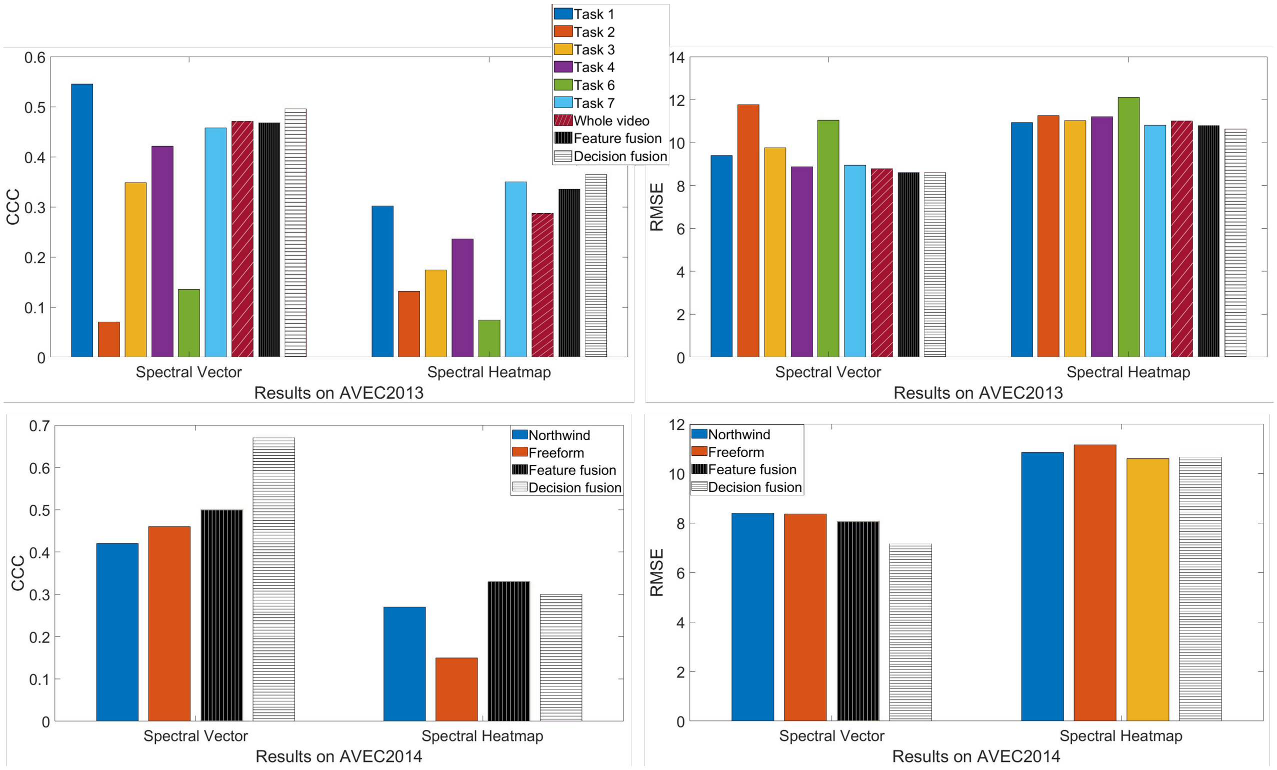


Fig. 9. Task-based results and fusion results achieved by two spectral representations on AVEC 2013 dataset and AVEC 2014 dataset.

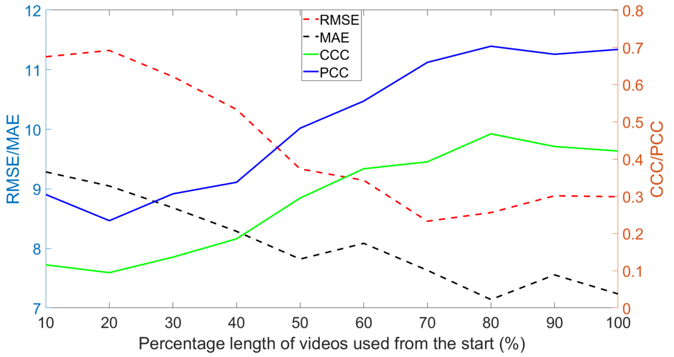


Fig. 10. Effect of video duration on depression estimation, using a varying percentage of the first frames of task 4.

them. In order to show the effectiveness/need for frequency alignment, we also report the average performance of models trained without any frequency alignment of the spectral features.

Table 3 and Table 4 compares the average (across all task-based and fusion experiments) performance of the two frequency alignment methods with the models where no alignment was done. The results on both the datasets, demonstrate that frequency alignment is necessary as both alignment methods achieved enhanced performance compared to models with no alignment. It can also be observed that the two frequency alignment methods achieved similar performance, with the proposed frequency selection method performing slightly better than zero-padding. Both methods have their own advantages: while zero-padding can increase the resolution of the spectral signals, the use of frequency selection prevents the original signal from distortion.

4.6.2 *Comparison of spectral representations*

We also compare the performance of the two spectral representations, i.e., spectral heatmap and spectral vector, described in Section 3.2.3. Fig.9, shows the fusion results as well as the results achieved by each task, individually. The performance from spectral vectors is significantly higher compared to the performance from the spectral heatmaps, across all experiments. One possible reason behind this outcome is that the number of the training examples (50 for training and 50 for validation) is too small to train CNNs (which usually have large number of trainable parameters) without overfitting. Secondly, we conducted feature selection before feeding spectral vectors to ANNs. This means that a large part of less informative behaviour information has been removed before the model training, which makes: 1. ANN models have smaller input layer and less parameters, making them easier to be trained by a small dataset; 2. the reduced data is more compact and has less noise.

4.7 Comparison to the state-of-the-art

We compare the performance of our top 2 best systems to state-of-the-art results on both datasets, in Table 5 and Table 6. For AVEC 2013 dataset, as only 39 test videos contain all tasks, it is not appropriate to compare any task-based result shown in Fig.9, to other published works. Instead, we report the results obtained by spectral representations extracted from whole videos of AVEC 2013 dataset. In particular, our best system in AVEC 2013 applied zero-padding to align frequencies and used spectral vector as the spectral representation. As shown in Table 5, our system achieved

Method	MAE	RMSE
Baseline [25]	10.88	13.61
Kachele et al. [68]	8.97	10.82
Meng et al. [29]	9.14	11.19
Cummins et al. [36]	N.A.	10.45
Wen et al. [17]	8.22	10.27
Kaya et al. [69]	7.86	9.72
Zhu et al. [19]	7.58	9.82
Ma et al. [70]	7.26	8.91
Jazaery et al. [42]	7.37	9.28
Melo et al. [28]	6.40	8.26
Zhou et al. [18]	6.20	8.28
Ours (Sel+SV)	6.67	8.11
Ours (Pad+SV)	**6.16**	**8.10**

TABLE 5
Comparison between our top 2 best systems and other works on **AVEC 2013** test set

Method	MAE	RMSE
Baseline [26]	8.86	10.86
Perez et al. [71]	9.35	11.91
Sidorov et al. [72]	11.20	13.87
Kaya et al. [73]	7.96	9.97
Zhu et al. [19]	7.47	9.55
Jazaery et al. [42]	7.22	9.20
Melo et al. [28]	6.59	8.31
Jan et al. [20]	6.68	8.01
Zhou et al. [18]	6.21	8.39
Ours(Sel+SV+Freeform)	6.78	8.30
Ours(Pad+SV+Freeform)	6.85	8.36
Ours(Pad+SV+Dec-fusion)	6.04	7.25
Ours(Sel+SV+Dec-fusion)	**5.95**	**7.15**

TABLE 6
Comparison between our top 2 best systems for single task/fusion and other works on **AVEC 2014** test set

the best RMSE and MAE results with 2.2% and 0.7% improvement respectively, over the current state-of-the-art. In terms of correlation metrics, the CCC and PCC of the second best system (Sel+SV) reached 0.68 and 0.75, respectively, while the best system (Pad+SV) achieved 0.60 and 0.73, respectively. The detailed predictions of the system (Sel+SV) is visualized in Fig. 11(a).

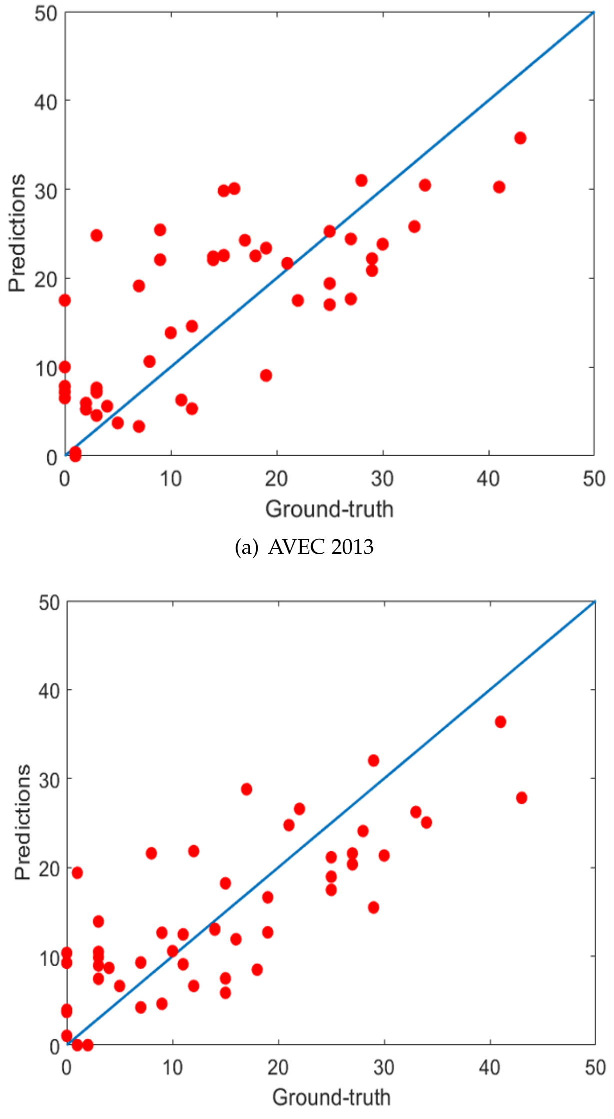
Both of our top 2 best systems in AVEC 2014 dataset outperform the current state-of-the-art. Our best system used selection to align frequencies for each task, i.e. Northwind and Freeform, and used spectral vectors as the representations. Meanwhile, the second best system used zero padding to align frequencies for each task and used spectral vectors as the representations. The final predictions of both systems were computed through decision-level fusion of the two tasks. As we can see from Table 6, our best system achieved RMSE and MAE results with 4.2% and 10.7% improvement compared to the current state-of-the-art. The CCC results of the best system (Sel+SV+Dec-fusion) and second best system (Pad+SV+Dec-fusion) are 0.67 and 0.63, respectively while the PCC results of both systems are 0.78. When only a single video set is used, our best results still outperformed all listed approaches (The result reported in [20] used both Northwind and FreeForm videos). The detailed predictions of our best system (Sel+SV+Dec-fusion) is visualized in Fig. 11(b).

5 Conclusion

This paper proposed a novel video-based automatic depression analysis approach using automatically detected facial behaviour primitives. As long-term temporal dynamics are important asset for depression analysis, the proposed approach first employs Fourier Transform to convert time-series behaviour signals to frequency domain as spectral signals, where each component in spectral signal encodes different frequency information of the whole video. As a result, the produced spectral signals contain multi-scale video-level temporal information. However, due to the variation in length of original videos, the length of their corresponding time-series behaviour signals and spectral signals are also variable. To allow spectral signals to be easily processed by standard ML models, we also propose two frequency alignment methods. Additionally, we also propose two spectral representations, i.e., spectral heatmap and spectral vector, to encode aligned spectral signals, allowing them to be learned by CNNs and ANNs, respectively.

We evaluated the proposed approach on AVEC 2013 and AVEC 2014 datasets, as the videos in each of them contain the same tasks. There are a series of studies conducted in our paper. Firstly, the analysis of facial behaviour primitives show that AU15, AU17, AU12, AU04, and AU09 are the most valuable behaviour primitives for depression estimation. Then, the task-based experiments and fusion experiments demonstrated that the contents of task affect depression estimation results significantly. Also, detecting depression from multiple tasks usually generate better results than using a single task alone. Thirdly, we compared two proposed frequency alignment methods, i.e, zero-padding and frequency selection. The results showed that they achieved similar results. Meanwhile, the comparison between the two spectral representations illustrated that spectral vectors clearly outperform the spectral heatmap. However, we believe that the performance of spectral heatmap can be potentially enhanced if more training data is available, as the amount of training data in the current audio-visual depression databases is not enough to train deep CNNs. Finally, we also compared our best systems to the state-of-the-art works. The results clearly showed that our approach outperform all other works.

As mentioned above, the performance of using CNN to train from spectral heatmaps can be potentially improved if more training data is available. Consequently, our future work will focus on collecting a large database for video-based depression analysis. Meanwhile, as this paper only used automatically detected AUs, gaze and head pose, for extracting frame-wise representation, which still ignores some other potentially useful information (e.g. microexpressions, speech, etc.). In future, we plan to explore what other kinds of behaviour primitives could be useful for automatic depression analysis.



(a) AVEC 2013

(b) AVEC 2014

Fig. 11. Predictions of our best systems on AVEC 2013 (top) and AVEC 2014 (bottom) datasets

Acknowledgments

The authors would like to thank the NIHR Nottingham Biomedical Research Centre (BRC) for its support. The work of Song was also supported by the Horizon Centre for Doctoral Training, School of Computer Science, University of Nottingham and Shenzhen University. The work of Shen is supported by National Natural Science Foundation of China under Grant No. 61672357,91959108 and the Science and Technology Funding of Guangdong Province (2018A050501014). The views represented are the views of the authors alone and do not necessarily represent the views of the Department of Health in England, NHS, or the National Institute for Health Research. Many thanks to Dr. Aaron Jackson for his continuous technical support.

References

[1] F. Edition, A. P. Association *et al.*, *Diagnostic and statistical manual of mental disorders*. Washington, American Psychological Association, 1994.

[2] Office for National Statistics (2017). Suicides in the UK: 2016 registrations. Available at: https://www.ons.gov.uk.

[3] N. Craddock and L. Mynors-Wallis, "Psychiatric diagnosis: impersonal, imperfect and important," *The British Journal of Psychiatry*, vol. 204, no. 2, pp. 93–95, 2014.

[4] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–7.

[5] "People with mental health problems still waiting over a year for talking treatments," https://www.mind.org.uk/news-campaigns/news/people-with-mental-health-problems-still-waiting-over-a-year-for-talking-treatments/, 2013, [Online; accessed 25-July-2019].

[6] H. Ellgring, *Non-verbal communication in depression*. Cambridge University Press, 2007.

[7] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[8] A. Stuhrmann, T. Suslow, and U. Dannlowski, "Facial emotion processing in major depression: a systematic review of neuroimaging findings," *Biology of mood & anxiety disorders*, vol. 1, no. 1, p. 10, 2011.

[9] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and vision computing*, vol. 32, no. 10, pp. 641–647, 2014.

[10] Y. E. Chentsova-Dutton, J. L. Tsai, and I. H. Gotlib, "Further evidence for the cultural norm hypothesis: Positive emotion in depressed and control european american and asian american women." *Cultural Diversity and Ethnic Minority Psychology*, vol. 16, no. 2, p. 284, 2010.

[11] J.-G. Gehricke and D. Shapiro, "Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion," *Psychiatry Research*, vol. 95, no. 2, pp. 157–167, 2000.

[12] B. Renneberg, K. Heyn, R. Gebhard, and S. Bachmann, "Facial expression of emotions in borderline personality disorder and depression," *Journal of behavior therapy and experimental psychiatry*, vol. 36, no. 3, pp. 183–196, 2005.

[13] D. M. Sloan, M. E. Strauss, and K. L. Wisner, "Diminished response to pleasant stimuli by depressed women." *Journal of abnormal psychology*, vol. 110, no. 3, p. 488, 2001.

[14] S. Scherer, G. Stratou, and L.-P. Morency, "Audiovisual behavior descriptors for depression assessment," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 135–140.

[15] I. B. Goldstein, "Role of muscle tension in personality theory." *Psychological Bulletin*, vol. 61, no. 6, p. 413, 1964.

[16] R. Rana, S. Latif, R. Gururajan, A. Gray, G. Mackenzie, G. Humphris, and J. Dunn, "Automated screening for distress: A perspective for the future," *European journal of cancer care*, p. e13033, 2019.

[17] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, 2015.

[18] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[19] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, 2017.

[20] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2018.

[21] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—hybrid architectures," *IEEE Transactions on Affective Computing*, 2018.

[22] J. L. Tsai, N. Pole, R. W. Levenson, and R. F. Muñoz, "The effects of depression on the emotional responses of spanish-speaking latinas." *Cultural Diversity and Ethnic Minority Psychology*, vol. 9, no. 1, p. 49, 2003.

[23] H.-U. Fisch, S. Frey, and H.-P. Hirsbrunner, "Analyzing nonverbal behavior in depression." *Journal of abnormal psychology*, vol. 92, no. 3, p. 307, 1983.

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

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

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

[27] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella *et al.*, "The distress analysis interview corpus of human and computer interviews." in *LREC*, 2014, pp. 3123–3128.

[28] W. C. de Melo, E. Granger, and A. Hadid, "Combining global and local convolutional 3d networks for detecting depression from facial expressions."

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

[30] T. Al Hanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," in *Proc. Interspeech*, 2018, pp. 1716–1720.

[31] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias *et al.*, "Depression assessment by fusing high and low level features from audio, video, and text," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 27–34.

[32] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 33–40.

[33] Y. Gong and C. Poellabauer, "Topic modeling based multi-modal depression detection," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 69–76.

[34] E. A. Stepanov, S. Lathuiliere, S. A. Chowdhury, A. Ghosh, R.-L. Vieriu, N. Sebe, and G. Riccardi, "Depression severity estimation from multiple modalities," in *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*. IEEE, 2018, pp. 1–6.

[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] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: a multimodal approach," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 11–20.

[37] V. Jain, J. L. Crowley, A. K. Dey, and A. Lux, "Depression estimation using audiovisual features and fisher vector encoding," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 87–91.

[38] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 255–259.

[39] K. Anis, H. Zakia, D. Mohamed, and C. Jeffrey, "Detecting depression severity by interpretable representations of motion dynamics," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 739–745.

[40] F. Hao, G. Pang, Y. Wu, Z. Pi, L. Xia, and G. Min, "Providing appropriate social support to prevention of depression for highly anxious sufferers," *IEEE Transactions on Computational Social Systems*, 2019.

[41] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding," *IEEE Transactions on Multimedia*, 2018.

[42] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Transactions on Affective Computing*, 2018.

[43] B. Sun, Y. Zhang, J. He, L. Yu, Q. Xu, D. Li, and Z. Wang, "A random forest regression method with selected-text feature for depression assessment," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 61–68.

[44] J. Rottenberg, K. L. Kasch, J. J. Gross, and I. H. Gotlib, "Sadness and amusement reactivity differentially predict concurrent and prospective functioning in major depressive disorder." *Emotion*, vol. 2, no. 2, p. 135, 2002.

[45] W. Gaebel and W. Wölwer, "Facial expressivity in the course of schizophrenia and depression," *European archives of psychiatry and clinical neuroscience*, vol. 254, no. 5, pp. 335–342, 2004.

[46] D. M. Sloan, M. E. Strauss, S. W. Quirk, and M. Sajatovic, "Subjective and expressive emotional responses in depression," *Journal of affective disorders*, vol. 46, no. 2, pp. 135–141, 1997.

[47] L. I. Reed, M. A. Sayette, and J. F. Cohn, "Impact of depression on response to comedy: A dynamic facial coding analysis." *Journal of abnormal psychology*, vol. 116, no. 4, p. 804, 2007.

[48] A. Z. Brozgold, J. C. Borod, C. C. Martin, L. H. Pick, M. Alpert, and J. Welkowitz, "Social functioning and facial emotional expression in neurological and psychiatric disorders," *Applied Neuropsychology*, vol. 5, no. 1, pp. 15–23, 1998.

[49] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[50] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. ACM, 2019, pp. 3–12.

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

[52] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 89–96.

[53] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE journal of biomedical and health informatics*, vol. 22, no. 2, pp. 525–536, 2018.

[54] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 41–48.

[55] M. Nasir, A. Jati, P. G. Shivakumar, S. Nallan Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 43–50.

[56] X. Ma, H. Yang, Q. Chen, D. Huang, and Y. Wang, "Depaudionet: An efficient deep model for audio based depression classification," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 35–42.

[57] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, "Measuring depression symptom severity from spoken language and 3d facial expressions," *arXiv preprint arXiv:1811.08592*, 2018.

[58] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[59] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 158–165.

[60] S. Jaiswal, S. Song, and M. Valstar, "Automatic prediction of depression and anxiety from behaviour and personality attributes,"

in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[61] S. Song, E. Sánchez-Lozano, M. Kumar Tellamekala, L. Shen, A. Johnston, and M. Valstar, "Dynamic facial models for video-based dimensional affect estimation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[62] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[63] F. Lindsten, "A remark on zero-padding for increased frequency resolution," *Sitio web: http://goo. gl/uBMFTw*, 2010.

[64] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 1578–1585.

[65] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[66] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[68] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," *depression*, vol. 1, no. 1, 2014.

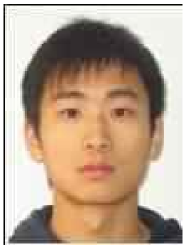
[69] H. Kaya and A. A. Salah, "Eyes whisper depression: A cca based multimodal approach," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 961–964.

[70] X. Ma, D. Huang, Y. Wang, and Y. Wang, "Cost-sensitive two-stage depression prediction using dynamic visual clues," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 338–351.

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

[72] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 81–86.

[73] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2014, pp. 19–26.



Siyang Song is a PhD student in the Computer Vision Lab and Horizon Center for Doctoral Training at the University of Nottingham. His external partner is Shenzhen University. His research interests include automatic emotion, personality and depression analysis using various Computer Vision techniques.

Shashank Jaiswal is a post-doctoral Research Fellow at the School of Computer Science, University of Nottingham. He received his PhD in computer science at the University of Nottingham in 2018. His research interests include automatic facial expression recognition and its applications in the diagnosis of mental health conditions.

Linlin Shen is currently a professor at Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. He is also a Honorary professor at School of Computer Science, University of Nottingham, UK. He serves as the director of Computer Vision Institute and China-UK joint research lab for visual information processing. He received the BSc and MEng degrees from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree from the University of Nottingham, Nottingham, U.K. He was a Research Fellow with the University of Nottingham, working on MRI brain image processing. His research interests include deep learning, facial recognition, analysis/synthesis and medical image processing. Prof. Shen is listed as the Most Cited Chinese Researcher by Elsevier. He received the Most Cited Paper Award from the journal of Image and Vision Computing. His cell classification algorithms were the winners of the International Contest on Pattern Recognition Techniques for Indirect Immunofluorescence Images held by ICIP 2013 and ICPR 2016.

Michel Valstar is an Associate Professor in the Computer Vision and Mixed Reality Labs at the University of Nottingham. He received his masters degree in Electrical Engineering at Delft University of Technology in 2005 and his PhD in computer science with the intelligent Behaviour Understanding Group (iBUG) at Imperial College London in 2008. His main interest is in automatic recognition of human behaviour. In 2011 he was the main organiser of the first facial expression recognition challenge, FERA 2011. In 2007 he won the BCS British Machine Intelligence Prize for part of his PhD work. He has published technical papers at authoritative conferences including CVPR, ICCV and SMC-B and his work has received popular press coverage in New Scientist and on BBC Radio.