Journal of Biomedical Informatics 83 (2018) 103–111

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automated depression analysis using convolutional neural networks from speech

Lang He[a,*], Cui Cao[b]

[a] NPU-VUB joint AVSP Research Lab, School of Computer Science, Northwestern Polytechnical University (NPU), Xi'an, China
[b] Moscow Institute of Arts, Weinan Normal University, Weinan, China

ARTICLE INFO

Keywords:
Depression
Automatic diagnosis
Median Robust extended Local Binary Patterns (MRELBP)
Speech processing

ABSTRACT

To help clinicians to efficiently diagnose the severity of a person's depression, the affective computing community and the artificial intelligence field have shown a growing interest in designing automated systems. The speech features have useful information for the diagnosis of depression. However, manually designing and domain knowledge are still important for the selection of the feature, which makes the process labor consuming and subjective. In recent years, deep-learned features based on neural networks have shown superior performance to hand-crafted features in various areas. In this paper, to overcome the difficulties mentioned above, we propose a combination of hand-crafted and deep-learned features which can effectively measure the severity of depression from speech. In the proposed method, Deep Convolutional Neural Networks (DCNN) are firstly built to learn deep-learned features from spectrograms and raw speech waveforms. Then we manually extract the state-of-the-art texture descriptors named median robust extended local binary patterns (MRELBP) from spectrograms. To capture the complementary information within the hand-crafted features and deep-learned features, we propose joint fine-tuning layers to combine the raw and spectrogram DCNN to boost the depression recognition performance. Moreover, to address the problems with small samples, a data augmentation method was proposed. Experiments conducted on AVEC2013 and AVEC2014 depression databases show that our approach is robust and effective for the diagnosis of depression when compared to state-of-the-art audio-based methods.

1. Introduction

Depression and anxiety disorders are highly prevalent worldwide, which have placed undue burden on individuals, families, and society. Studies suggest that effective treatments for depression can be aided by the detection of the problems at its early stages. According to the World Health Organization (WHO), depression will become the fourth most mental disorder by 2020 [1].

Depression is often difficult to diagnose because it manifests itself in different ways. The assessment methodologies for its diagnosis rely on subjective patient self-report or clinical judgments of symptom severity [2,3]. The Hamilton Rating Scale for Depression (HAMD) [4] is currently the standard for depression severity estimation. It is worth noting that, evaluations by clinicians vary depending on their expertise and the used diagnosis methods, such as Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) [5], the Quick Inventory of Depressive Symptoms-Self Report (QIDS) [6], the Beck Depression Inventory (BDI) [7], the 10-item Montgomery–Asberg Depression Rating Scale (MADRS) [8], the 9–item Patient Health Questionnaire (PHQ–9) [9], and the PHQ–8 [10].

In recent years, some machine learning methods have been proposed utilizing audio cues for depression analysis [11–16]. Meanwhile, there is a wealth of research, which suggests that voice patterns have a close relationship with emotion and stress [17–19]. In [20], the author suggested that the analysis of voice patterns can be divided into three primary categories, including prosodics, the vocal tract, and the glottal source. Although hand-crafted features have been proven to obtain better performance for estimating depression severity. However, there are some limitations of handcrafted features for depression scale prediction. First, to design hand-crafted features requires a lot of effort (i.e., domain knowledge, labor and time, etc.). For example, Mel Frequency Cepstral Coefficients (MFCCs) are widely used in automatic speech and speaker recognition tasks. However, if we designed hand-crafted features like MFCCs, we should have task-specific knowledge of depression and to acquire such knowledge is time-consuming. Second, hand-crafted features may lose some useful information related to depression patterns. Specifically, some patterns of depression implied in the audiovisual signals cannot be well mined. Moreover, the concept of

* Corresponding author.
E-mail address: langhe@mail.nwpu.edu.cn (L. He).

https://doi.org/10.1016/j.jbi.2018.05.007
Received 6 November 2017; Received in revised form 25 April 2018; Accepted 12 May 2018
Available online 29 May 2018
1532-0464/

the designed features relies on people's subjective assumptions. Finally, it is difficult to select an appropriate toolkit to extract the features. Various available toolkits are widely used to extract low-level features, such as openSMILE [21], COVAREP [22], SPTK [23], KALDI [24], YAAFE [25], and OpenEAR [26]. Each existing toolbox is generally the result of a single laboratory's work. Different researchers considered features from their own perspective. There is no unified standard defining which feature is most useful for depression analysis.

Recently, deep learning has been successfully applied to various communities. Both theories and experiments have shown that deep learning can learn a lot of valuable information from the audiovisual signals. The deep learning method has several variants, such as single Layer Learning models, Probabilistic Models, Auto-Encoders and Convolutional Neural Networks. A more in-depth understanding of the deep learning methods the reader is referred to [27]. Among these different deep learning representations, Convolutional Neural Networks has been widely used to achieve state-of-the-art performance in many communities [28–30]. Moreover, it has been proved fairly efficient in texture classification scenario. In [31], the authors proved that the CNN-based method matched the state-of-the-art for the dataset with macroscopic images, and outperformed the best-published results on the microscopic images. The performance of proposed CNN architecture also surpass exist texture descriptors for forest species recognition. To the best of our knowledge, deep-learned features from spectrogram for depression recognition has not yet been explored. Accordingly, in this work we explore how the depression severity prediction can benefit from the adoption of CNN in learning spectrogram patterns of the speech.

From the machine learning perspective, depression analysis can be considered as a regression or classification problem (e.g., in AVEC2013 [14] and AVEC2014 [15] depression sub-challenges). Our goal is to predict the depression score called Beck Depression Inventory–II (BDI–II) of a subject from recorded audio.

In summary, the main contributions of this work can be summarized as follows. First, we develop an automated framework, which can effectively capture the vocal information for measuring the depression severity. Second, we find that complementary characteristics is existed between hand-crafted features and deep learned features for estimating the depression severity. Third, we propose a combination of the hand-crafted and the deep-learned features to effectively measure the severity of depression from speech. Finally, to address the problems with small samples, a data augmentation method was proposed. To the best of our knowledge, in our proposed approach, it is the first time that the deep learning technology is employed for depression diagnosis.

The remainder of this paper is organized as follows. Section 2 briefly discusses previous works on audio-based depression analysis and recognition. Section 3 provides more implementation details about the proposed framework. Section 4 introduces the dataset and experimental results. Conclusions and future challenges are discussed in Section 5.

2. Related work

Various depression recognition approaches have been proposed in the Depression Recognition Sub-Challenge (DSC) of the Audio-Visual Emotion Challenge and Workshop (AVEC2013, AVEC2014, AVEC2016 [32], AVEC2017 [33]).

Regression methods have been developed using the AVEC2013 and AVEC2014 data sets, and classification approaches considered the AVEC2016 and AVEC2017 data. In this work, we make use of the AVEC2013 and AVEC2014 data sets. Detailed description of the database can be referred to Sections 4.1 and 4.2. In our research, we focus on the recorded audio for the diagnosis of depression. In the following section, we briefly describe the competitive audio-based methods for measuring the depression severity.[1]

For AVEC2013 depression recognition [14], researchers have used audio baseline features extracted by using the freely available open-source Emotion and Affect Recognition (openEAR) [26] toolkit's feature extraction backend openSMILE [21]. The audio feature set consists of 2268 features, including 32 energy and spectral related low-level descriptors (LLD) × 42 functionals, 6 voicing related LLD × 32 functionals, 32 delta coefficients of the energy/spectral LLD × 19 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced/unvoiced durational features. In order to capture the dynamic, long-range characteristics, the authors segment the audio clips with fixed length segments (3 s), which shift at one second. Finally, Support Vector Regression (SVR) is used for learning and predicting.

In the AVEC2013 depression challenge, Williamson et al. [13] adopted the combination of eigenvalue spectra and coordination features to analyze the relationship between the vocal behaviors and the depression scales. With the coordination- and phoneme-rate-based features, they designed a Gaussian staircase regression system to predict the BDI–II scores for each audio data. PCA is also used for dimension reduction. Finally, the authors provided the minimum performance on the test sets with root mean square error (RMSE) of 7.42 and mean absolute error (MAE) of 5.75.

In [34], Moore et al. explored prosodics, the vocal tract, and parameters extracted directly from the glottal waveform to discriminate the depressed speech. They extracted about 200 prosodics, vocal tract, and glottal waveform measures from the depression database and translated them into 2000 statistics for study.

In [35], Nicholas et al. provided a comprehensive and exhaustive conclusion about the assessment and diagnosis of the depression and the suicide. They reviewed the important characteristics of paralinguistic speech affected by depression and suicide. They analyzed the patterns which were used in classification and regression issues. Finally, they provided an in-depth discussion about the current limitations and challenges.

In [36,16], the authors investigated the relation between vocal prosody and change in depression severity over time. They presented three hypotheses: (1) Naive listeners can distinguish the depressed participants and health controls from vocal recordings; (2) the quantitative features of vocal prosody can capture changes from the diagnosis of the depression; and (3) interpersonal relationships can also occurred in the severity of depression estimation procedure. Finally, they validated the hypotheses by experiments. The results showed that the analysis of vocal prosody is a valuable tool for depression analysis.

In [37–47], all of them use the audio feature provided by the AVEC2013 depression sub-challenge. In [48], they also explored a number of features, (1) estimated articulatory trajectories during speech production, (2) acoustic characteristics, and (3) acoustic-phonetic characteristics and (4) prosodic features. They are used and compared with different models to predict the Beck depression rating scale, such as support vector regression (SVR), a Gaussian backend, and decision trees.

In [49], Williamson et al. explored the interrelationships and complementary characteristics by extracting features from the speech source, system, and prosody. They fused the different feature domain to obtain a better performance. Finally, they combined Gaussian staircase regression with Extreme Learning Machine (ELM) classifiers, and get a test RMSE of 8.12.

For the AVEC2016 [32] and AVEC2017 [33] depression sub-challenge, the organizers provided the audio, video, and transcript files, but did not provide the original video clips. For the audio features of both

[1] Some of following works also used the video cues, while we only focus on the audio cues.

in AVEC2016 and AVEC2017, they used COVAREP (v1.3.2), a freely available open source Matlab toolbox for speech analyses [22]. Prosodic, voice quality, and spectral features were extracted by the COVAREP toolkit from the audio signals. In [50–52], all of them used the audio features provided by the AVEC2016 organizers. However, the baseline audio features does not include all of the features considered as useful for depression prediction (e.g., jitter, shimmer, etc.). Therefore, the authors in [53–55], also extracted another useful audio feature for depression recognition. In [53], they extracted spectral features, lower vocal tract physiology features, and loudness variation features, obtaining relatively better results for depression prediction. In [54], they extracted extended spectral and prosodic features, teager energy cepstral coefficients, session-level acoustic features, and phoneme-based features. They obtained F1 scores of 0.63 and 0.89 for depressed and not depressed classes respectively.

From the literature review we can see that hand-crafted audio features have limitations in diagnosing depression. Specially, hand-craft audio features are extracted by different toolboxes from the perspective of different researchers. To overcome these limitations, we explore a more robust representation for depression analysis, which could better capture valuable information from the vocal cues. That is to say, we propose a new approach based on the deep learning networks, for automatic estimation the severity of the depression scale.

3. Methodology

Feature design or feature extraction plays an important role in depression analysis tasks. In this work we combine hand- crafted features with deep-learned features for estimating the severity of depression. First, for hand-crafted features, we extract the Low Level Descriptors (LLD) from the raw audio clips and Median Robust extended Local Binary Patterns (MRELBP) features from the spectrograms of audio. Second, we use DCNN to directly learn the deep-learned features from the raw audio and spectrogram images. Finally, we describe the proposed joint fine-tuning method to combine the four streams for the final depression prediction. The proposed framework for automatic depression recognition is given in Fig. 1.

3.1. Hand crafted based feature extraction

For hand-crafted features, two different kinds of descriptors were adopted. The first one is the Median Robust Extended Local Binary Patterns (MRELBP), a novel descriptor for texture classification [56]. However, its application for depression recognition has yet not been explored. In this work, we apply MRELBP on spectrogram to extract textural features. The other one is the audio features extracted by openSMILE toolkit.

Table 1
38 low-level descriptors.

Energy&Spectral (32)
loudness (auditory model based), zero crossing rate, energy in bands from 250–650 Hz, 1 kHz-4 kHz, 25%, 50%, 75%, and 90% spectral roll-off points spectral flux, entropy, variance, skewness, kurtosis, psychoacousitc sharpness, harmonicity, flatness, MFCC 1–16
Voicing related (6)
F0 (sub-harmonic summation, followed by Viterbi smoothing), probability of voicing, jitter, shimmer (local), jitter (delta: "jitter of jitter"), logarithmic Harmonics-to-Noise Ratio (logHNR)

3.1.1. Audio features

The 2268 baseline audio features of AVEC2013 [14] and AVEC2014 [15] adopt the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) containing functionals on the 38 low-level descriptors (LLDs), which are extracted with the openSMILE toolkit [21]. These LLDs cover the spectral, cepstral, prosodic and voice quality information. 38 low-level descriptors (LLDs) are shown in Table 1.

To capture the valuable pattern of the depression, we try different strategies to segment the audio features. In our experiment, we use overlapping fixed length segments shifting forward at a rate of 1 s, while the size of the windows is 20 s, which can capture slow changing, long range characteristics.

Details for functionals can be found in [14]. For the audio features, 42 functionals on 32 energy and spectral related low-level descriptors (LLD), 32 functionals on 6 voicing related LLD, 19 functionals on 6 delta coefficients of the voicing related LLD, and 10 voiced/unvoiced durational features, resulting in 2268 feature vectors.

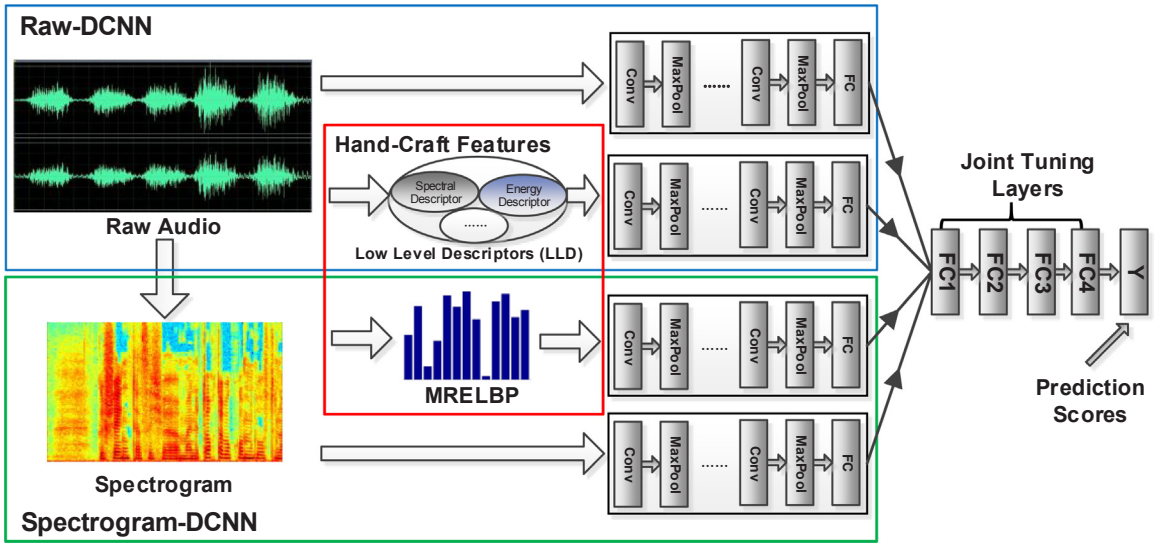

Fig. 1. Illustration of the proposed method for depression recognition using deep neural networks. The Raw-DCNN (Top) takes raw audio signals and low level descriptors (LLD) as input, while the Spectrogram-DCNN (Bottom) uses texture features as input. The red box in Fig. 1 is Hand-Crafted features. Other two arrows are Deep-Learned features. The predicted depression score is computed by aggregating or averaging the individual predictions per frame from four DCNNs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

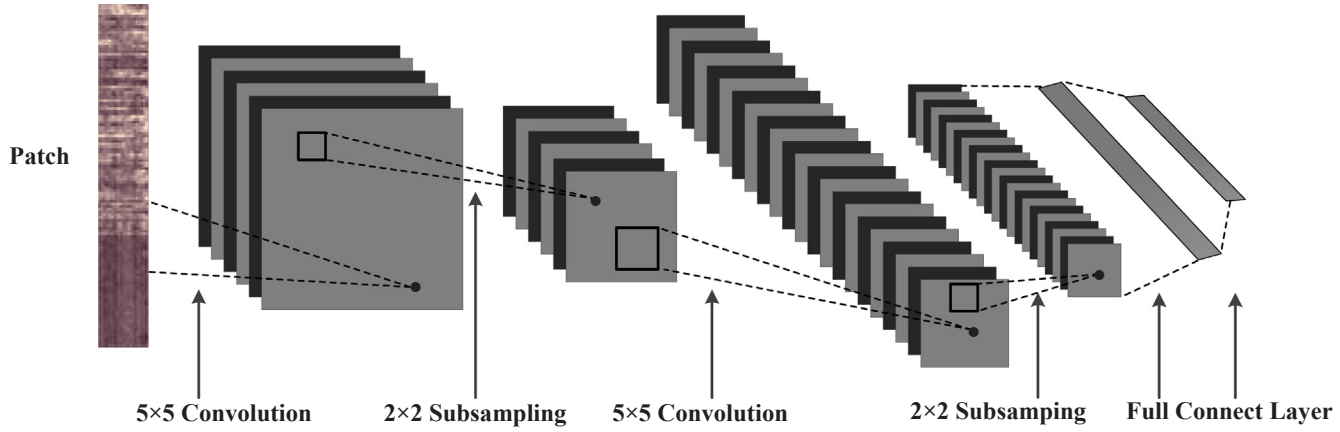

Fig. 2. The deep Convolutional Neural Network architecture.

3.1.2. Median Robust Extended Local Binary Patterns (MRELBP)

The LBP operator characterizes the spatial structure of a local image patch by encoding the differences between the pixel value of the central point and those of its neighbors, considering only the signs to form a binary pattern. Formally, given a pixel x_c in the image, the basic LBP response is calculated by comparing its value with those of its P neighboring pixels $\{x_{R,P,n}\}_{n=0}^{P-1}$, evenly distributed on a circle of radius R centered on x_c:

$$LBP_{R,P}(x_c) = \sum_{n=0}^{P-1} s(x_{R,P,n} - x_c)2^n$$
$$s(x) = \begin{cases} 1, & \text{if } x \geqslant 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (1)$$

Recently, Liu et al. [56] proposed the Median Robust Extended Local Binary Pattern (MRELBP), where the individual pixel intensities in Eq. (1) were replaced by a median filter response $\phi()$ to maximize the robustness of the representation to noise. Different from the traditional LBP, MRELBP compared regional image medians rather than raw image intensities, which can capture both microstructure and macrostructure texture information. For a more detail understanding of the methods, we refer the reader to [56].

3.2. Deep learning based features extraction

As DCNN has shown its advantages in learning the patterns of feature, we adopt it to learn the valuable characteristic information implied in the audiovisual signals. Our deep-learned features are extracted by using two different models. The first deep network extracts deep-learned audio features from frame-level raw waveforms, while the other deep network model directly learns feature representations from spectrogram images. We describe our methods in detail below.

3.2.1. Deep learned audio features

For the deep-learned audio features, we feed the frame-level raw waveform into the first CNN convolutional layer to learn a filter-bank representation which was equivalent to filter kernels in a time-frequency representation. In this method, if the raw waveform is filtered by the first strided convolutional layer, the output feature map will have the same as the spectrogram. Specifically, the parameters of the first convolutional layer (i.e., stride, filter length, and number of filters) corresponded to the parameters of spectrogram (i.e., mel-size, window size, and number of mel-bands, respectively).

3.2.2. Deep learned texture features

In this section, we describe the details for the deep-learned texture features, which were different from the deep-learned audio features.

Even as a commonly used neural network technology, CNN has its own limitations. First, CNN cannot process the high-resolution images. Second, it requires a lot of samples for training. Specifically, CNN can learn a large number of parameters through the training procedure based on the amount of the training data. To overcome the limitations mentioned above, we first segment the audio clips with different size. Several authors studied the appropriate length of segments for extracting reliable audio features. In [14,15] the authors proposed 20s to capture slow changing and long range characteristics. For the extraction of vocal patterns using CNN, we first conducted experiments using segment of 6s, and 20s. Then we proposed a data augmentation method to tackle with the small samples of the training data. First, Δ and ΔΔ features were extracted from the frequency domain of spectrograms. Second, following the above-mentioned step, the whole spectrograms image sequences were horizontally flipped. We rotated each image by each angle in -15°, -10°, -5°, 5°, 10°, 15°. Finally, we receptively obtained 14 times more data than the original images, Δ, ΔΔ: original images (1), flipped images (1), rotated images with six angles, and their flipped versions (12). In total, we obtained 42 times more data samples to train the model. After the above augmentation process, this makes the model robust for learning a lot of parameters of the input images.

The DCNN architecture used in our work has been proved effective to perform well on other tasks such as object recognition, action recognition, etc. It repeatedly adopts convolutional layers with 64 filters followed by max-pooling layers, inspired by [31]. The architecture is illustrated in Fig. 2. To improve the computational efficiency and boost the recognition accuracy, we resize the spectrogram image into 128×128.

In the following, we describe the CNN architecture with parameters. In our work, the input image size is 128×128. The convolutional layers have trainable filters (feature maps), which were applied across the entire image. The definition of the layers consisted of the filter size and stride, which was the distance between the applications of the filters. If the stride size is smaller than the filter size, the overlapped windows can be adopted for the filters. To learn the optimal hyperparameters, we conducted several experiments and obtained a filter size of 5×5 with stride size of 1.

For the pooling layers, we aimed to implement a non-linear down sampling function for dimensionality reduction and thus achieve translation invariance. In our study, we used different kernels and strides to carry out experiment. We found that the window size 2×2 with stride 2 get the best performance. Similar to [31], the fully-connected layers connect, all the neurons of one layer with that of the next one. In our work, depression severity measurement is considered as a regression problem from the point of machine learning view. Therefore, Euclidean loss was used as the loss function of the network, which was suitable for regression. Mathematically, the Euclidean loss function E computes the sum of squared differences of its two inputs, which can be

Table 2
The number of samples of training, development, and test set on the AVEC2014 database.

Partitions	Number of Samples
Train	100
Dev.	100
Test	100

written as:

$$E = \frac{1}{2N}\sum_{i=1}^{N} \|\hat{y}_i - y_i\|^2 \quad (2)$$

where N is the number of samples, $\hat{y}_i$ denotes the output from the network, and y_i represents the ground truth (BDI–II score).

3.3. Joint fine-tuning method

In our approach, the Raw-DCNN and the Spectrogram-DCNN are able to predict BDI–II scores separately. To capture the complementary information within the two used models, we propose joint fine-tuning method to boost the recognition performance. Specifically, four fully connected layers are concatenated as feature layers in both the raw and spectrogram networks. Euclidean loss function is still used for regression in our task. In the training process, the four DCNNs are trained separately, and then the joint fine-tuning is created using the architecture with joint tuning layers, as shown in Fig. 1. Meanwhile, the dropout method is adopted for reducing over-fitting.

4. Experimental evaluation

In this section, the datasets used for the experiments are introduced firstly in Sections 4.1 and 4.2. In Section 4.3, we briefly detail the experimental setup. Finally, the experimental results are provided in Section 4.4.

4.1. AVEC2013 depression database

AVEC2013 [14] used a subset of the audio-visual depressive language corpus (AVDLC), which included 340 video clips of 292 subjects, performing a Human–Computer Interaction task while being recorded by a webcam and a microphone. The AVEC2013 database consist of 14 different tasks which were Power Point guided: e.g., sustained vowel phonation, sustained loud vowel phonation, and sustained smiling vowel phonation; speaking out loud while solving a task; Counting from 1 to 10, etc. The subjects were recorded between one and four times, with a period of two weeks between the measurements. The average age was 31.5 years with a range of 18–63 years. The length of each recording varied from 20 to 50 min, with an average duration of 25 min per recording. The 16-bit audio was recorded at a sampling rate of 41 KHz. The videos, with frames of 640 × 480 pixels and 24-bits per pixel, were recorded at 30 frames per second. For the depression sub-challenge, there were 150 videos from 82 subjects. The recordings were split into three partitions: a training, development, and test set of 50 recordings each.

The depression levels were labeled per clip using Beck Depression Inventory-II (BDI-II). Final BDI-II scores range from 0 to 63 (0–13 no or minimal depression; 14–19 mild depression; 20–28 moderate depression; 29–63 severe depression).

4.2. AVEC2014 depression database

AVEC2014 corpus [15] was a subset of the AVEC2013 corpus. The AVEC2014 corpus consisted of recordings of 2 different human–computer interaction tasks. Each of the tasks was supplied as separate recordings. In total, the corpus includes 300 videos with the duration ranging from 6 s to 4 min. The two tasks included a reading task and a spontaneous speech task, which are described below:

- Northwind - Participants read aloud an excerpt of the fable “Die Sonne und der Wind” (The North Wind and the Sun). (German).
- Freeform - Participants respond to one of a number of questions such as: “What is your favorite dish?” or discuss a sad childhood memory (German).

Each recording was labeled with BDI-II severity of depression. For AVEC 2014 depression sub-challenge, every task was split into three partitions: a training, development, and test set of 50 recordings each. In our experiments, we combined the training sets of the Northwind dataset and Freeform dataset to train the models, the development sets to verify the performance, and the test sets to test the models. Therefore, the training, development, and test set have 100 recordings, respectively (Table 2).

4.3. Experimental setup and evaluation measures

In this sub-section, we describe the experimental setup and evaluation measures in detail.

4.3.1. Experimental setup

As mentioned above, we use the Raw-DCNN and Spectrogram-DCNN to measure the severity of depression. As shown in Fig. 1, each part is implemented using a DCNN architecture.

The dataset haven’t included the spectrograms. We first segment the audio with 6s length segments shift forward at a rate of one second as the augmented samples. After the segmentation, we convert each audio segment into mono by calculating the mean of the left and right channels, and then we normalize the data by mapping row minimum and maximum values between -1 and 1. In other words, the normalization process is to restrict the amplitude ranges. For the audio data, sampling frequency is resampled to 16 kHz. In our work, to make the input data smaller, we performed the discrete Fourier transform (DFT) to obtain a time-frequency representation of the audio. For DFT parameter setup, we adopt a Hanning window function of 23 ms and 50% overlap. After the above steps, a spectrogram is generated for every audio clip. For the length of LLD features, we tried various segment lengths and found 20 s length as optimum. The 20 s of segment can capture both slow changing and long range characteristics. For estimating the LBP feature and the Median Robust extended Local Binary Patterns (MRELBP) feature, the parameters have been selected to obtain the best performance for texture classification. In this work, we consider (i) 2 radii R values, $R = 1$ and $R = 3$, and (ii) number of local neighboring points set as $P = 8$. Like most other LBP variants, we also use the uniform encoding scheme [57] for LBP and MRELBP. For MRELBP, the authors [56] have proved that the uniform encoding scheme can obtain the striking texture classification accuracy.

For the DCNN architecture, the networks are trained with stochastic gradient using caffe deep learning toolbox [58] with a batch size 32. For both of the Raw-DCNN and Spectrogram-DCNN, the training procedure starts from scratch. Euclidean loss is considered as the loss function for regression. The number of iterations for Raw model and Spectrogram model were set to 200,000 and 400,000, respectively. The parameters of the two deep networks are selected by experience and followed recommendation in another works[59]. The learning rate was set to 10^{-3} reduced by polynomial with gamma equals to 0.5. The momentum was set to 0.9 with weight decay equals to 0.0002. All experiments were conducted using NVIDIA Quadro K2200 with 4G memory.

In our experiments, the joint tuning layers are designed as two fully connected layers with 512 and 256 hidden units, respectively. To use the advantage of the deep-learned model and hand-crafted model, we

proposed an integration method for Raw-DCNN and Spectrogram-DCNN using a joint fine-tuning method, which achieves better results than the two models. In the joint fine-tuning procedure, we retrained the top layers, and freezed other layers of the two trained networks.

4.3.2. Evaluation metric

The depression severity recognition performance is assessed in terms of mean absolute error (MAE) and root mean square error (RMSE) between the prediction and reported BDI–II values.

The MAE was computed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \tilde{y}_i| \tag{3}$$

And the RMSE was computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2} \tag{4}$$

where N denotes the number of data samples, y_i is the ground truth and $\tilde{y}_i$ represents the predicted value of i-th sample, respectively.

4.4. Experimental results

In the following, we first compare the performance of LBP feature with the MRELBP feature. Then, we compare the performance of hand-crafted features with that of deep-learned features for depression scale prediction. Finally, we compare our results to the ones from other state-of-art methods.

4.4.1. Performance of single models

The performances of depression recognition on AVEC2013 and AVEC2014 databases are shown in Tables 3 and 4, respectively. In our work, we first described the results using single models without any joint tuning procedure. Table 3 shown that the deep-learned features obtained the better results with MAE 8.4832 and RMSE 10.4561 on the test set. In comparison with the performance of AVEC2013, AVEC2014 obtains better results. As shown in Table 4, the deep-learned features also obtained the better performances with MAE 8.6014 and RMSE 10.4413 on the test set. These results showed that the deep learned model was important for depression severity prediction, and the spectrogram DCNN can represent the characteristics of depression. Moreover, the deep learned model could reduce some effort to design and find the suitable hand-crafted features for depression scale prediction.

4.4.2. Overall performance by fusing the individual models

In our experiments, to capture the complementary information with the deep-learned features using DCNN and the hand-crafted features, we calculate the performance by fusing the hand-crafted features and deep-learned features. The results are shown in the first row in Tables 5 and 6 for AVEC2013 and AVEC2014, respectively. It can be seen from the table that, when averaging is adopted, the RMSE and MAE obtained are 10.2261 and 8.2323 on the AVEC2013 database, respectively. On AVEC2014, the MAE of 8.2204 and RMSE of 10.1284 are obtained. The results showed that by fusing the hand-crafted and the deep-learned model, the overall performance can be improved than adopting the single model which means the necessity of using both hand-crafted and deep-learned features for depression scale prediction.

Table 3
Performance of hand-crafted and deep-learned features on the development set and test set of AVEC2013.

Partition	Methods		RMSE	MAE
Dev.	Hand crafted model	LBP	9.3507	7.7314
		MRELBP	9.1673	7.5455
		LLD	9.3154	7.6502
	Deep learned model	Waveform	9.3896	7.8184
		Spectrogram	9.1129	7.5371
Test	Hand crafted model	LBP	10.9312	9.2443
		MRELBP	10.5611	8.6580
		LLD	10.6418	8.8935
	Deep learned model	Waveform	11.0983	9.4484
		Spectrogram	10.4561	8.4832

Table 4
Performance of hand-crafted and deep-learned features on the development set and test set of AVEC2014.

Partition	Methods		RMSE	MAE
Dev.	Hand crafted model	LBP	9.3478	7.5699
		MRELBP	9.1523	7.5026
		LLD	9.3000	7.5514
	Deep learned model	Waveform	9.3770	7.8813
		Spectrogram	9.1100	7.4969
Test	Hand crafted model	LBP	10.8211	8.7489
		MRELBP	10.4618	8.6420
		LLD	10.5648	8.6800
	Deep learned model	Waveform	10.9014	8.7810
		Spectrogram	10.4413	8.6014

Table 5
Overall performance on the development set and test set of AVEC2013.

Partition	Methods	RMSE	MAE
Dev.	Hand & Deep (Ave.)	9.1001	7.4456
	Hand & Deep (Joint Tuning)	9.0000	7.4210
Test	Hand & Deep (Ave.)	10.2261	8.2323
	Hand & Deep (Joint Tuning)	10.0012	8.2012

Table 6
Overall performance on the development set and test set of AVEC2014.

Partition	Methods	RMSE	MAE
Dev.	Hand & Deep (Ave.)	9.0089	7.4213
	Hand & Deep (Joint Tuning)	9.0001	7.4211
Test	Hand & Deep (Ave.)	10.1284	8.2204
	Hand & Deep (Joint Tuning)	9.9998	8.1919

4.4.3. Overall performance by joint tuning

In our research, we also conducted the experiments by using joint tuning method. The results are shown in the last row in Tables 5 and 6 on AVEC2013 and AVEC2014, respectively. Table 5 illustrates that the results after joint tuning the models were MAE 8.2012 and RMSE 10.0012 on the AVEC2013 database. While on the AVEC2014 database,

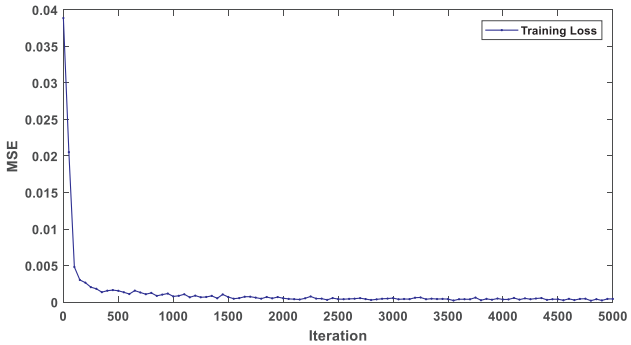

Fig. 3. The training MSE loss decreases and converges during joint fine-tuning for final depression scale prediction.

Table 7
Overall performance on the development set and test set (AVEC2013 + AVEC2014).

Partition	Methods	RMSE	MAE
Dev.	Hand & Deep (Ave.)	8.9971	7.4200
	Hand & Deep (Joint Tuning)	8.8920	7.4118
Test	Hand & Deep (Ave.)	10.0009	8.2323
	Hand & Deep (Joint Tuning)	9.8874	8.1901

Table 8
AVEC2013 - Comparison to state-of-the-art results. Note that the listed results use audio data only.

Partition	Methods	RMSE	MAE
Dev.	Baseline [14]	10.75	8.66
	Meng et al. [38]	8.82	7.09
	Williamson et al. [13]	N/A	N/A
	Ours	***9.0000***	***7.4210***
Test	Baseline [14]	14.12	10.35
	Meng et al. [38]	11.19	9.14
	Williamson et al. [13]	7.42	5.75
	Ours	***10.0012***	***8.2012***

as shown in Table 6, the best results were MAE of 8.1919, and RMSE of 9.9998. The results implied that the proposed joint tuning method performance was improved when employing both the handcrafted and deep learned models. During the training process, the performance is measured by Euclidean loss in the joint fine-tuning process. Fig. 3 shows convergences of the MSE loss during joint fine-tuning for final depression scale prediction.

In addition, we also combine the AVEC2013 and AVEC2014 as a single database to predict the depression scale. As shown in Table 7, the results after combining the two databases are the MAE of 8.1901 and the RMSE of 9.8874. A potential reason for this is: the new enlarged database has more data samples for training and the DCNN models can better predict the depression scores.

4.4.4. Comparison with previous works

In Tables 8 and 9, we compare our depression recognition results. Using the proposed approach, we combined hand-crafted features and deep-learned features, to state-of-the-art results using other audio features, for the AVEC2013 and AVEC2014 databases, respectively. The indicated results were similar to those reported by previously cited

Table 9
AVEC2014 - comparison to state-of-the-art results. Note that the listed results use audio data only.

Partition	Methods	RMSE	MAE
Dev.	Baseline [15]	11.52	8.93
	Jain et al. [41]	11.51	9.75
	Jan et al. [40]	10.69	8.92
	Senoussaoui et al. [46]	10.09	7.41
	Parez et al. [43]	9.79	7.75
	Kachele et al. [45]	N/A	N/A
	Mitra et al. [48]	7.71	6.10
	Ours	***9.0001***	***7.4211***
Test	Baseline [15]	12.567	10.036
	Jain et al. [41]	10.25	8.40
	Jan et al. [40]	11.30	9.10
	Senoussaoui et al. [46]	12.71	9.82
	Parez et al. [43]	11.92	9.36
	Kachele et al. [45]	9.18	7.10
	Mitra et al. [48]	11.10	8.83
	Ours	***9.9998***	***8.1919***

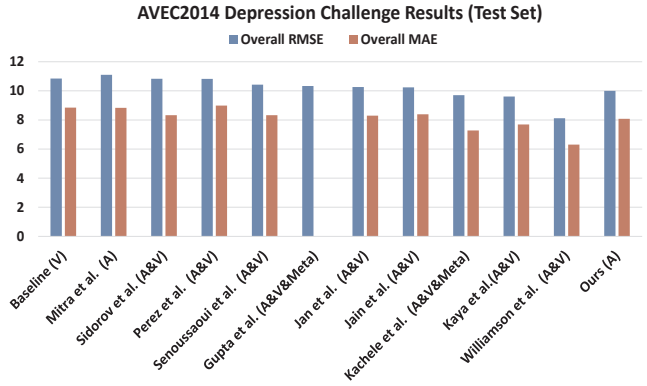

Fig. 4. AVEC2014 - Comparison with techniques of depression recognition using audio (A) and visual (V) features.

studies. We should note here that these results have also been obtained on the combined dataset of Freeform and Northwind. As shown in Tables 8 and 9, our approach provided better results than most of the state-of-the-art research that has been conducted. To make a fair comparison with other works, we use the training, validation, and test set provided by the database providers. The new augmented samples are generated from the original training, development and testing set. In our work, we applied the same data augmentation approach on the three datasets.

As shown in Table 8, it is clearly demonstrated that the proposed method outperforms all the other methods but one [13]. In [13], the authors adopted a feature space to capture useful information based on the eigenvalue spectra - coordination features - and combined them with a feature set involving average phonetic durations, i.e., phonetic-based speaking rates. While in Table 8, the proposed method surpasses all the methods except one [45]. In [45], Kachele et al. propose an approach based on abstract meta information about individual subjects and also prototypical task and label dependent templates to infer the respective emotional states. They obtained better results in the depression challenge task.

In Fig. 4, we report our results on the AVEC2014 dataset compared to reported state-of-the-art results using both audio (A) and video (V) features. As they can be seen using only audio features, our methods provided comparable results to multi-modal approaches for depression recognition.

5. Conclusions and future works

Depression is a serious psychological disorder. Computer aided technologies have been investigated to assist psychologists in the assessment of depression levels. To improve the accuracy of automatic depression recognition from speech signals, we proposed a new method based on deep learning and traditional method, which we employed to overcome the difficulties caused by designing hand-crafted features for depression recognition. In the proposed method, we use the raw and spectrogram DCNN to model the characteristic information of depression. Moreover, we also proposed to adopt joint tuning layers, to combine the raw and spectrogram DCNN, which can improve the performance of depression recognition. Experimental results on two depression dataset, AVEC2013 and AVEC2014, have demonstrated that our approach obtain superior performance compared with other audio-based methods for depression recognition. In our future work, we will explore more powerful regression models to further improve the accuracy of depression recognition.

Conflict of interest

No potential conflict of interest relevant to this article was reported.

Acknowledgment

This work is supported by the Shaanxi Provincial International Science and Technology Collaboration Project (grant 2017KW-ZD-14), the National Natural Science Foundation of China (grant 61273265), the VUB Interdisciplinary Research Program through the EMO-App project, and the program of China Scholarship Council (CSC) (No. 201606290171).

References

[1] C. Mathers, D.M. Fat, J.T. Boerma, The Global Burden of Disease: 2004 Update, World Health Organization, 2008.
[2] A.T. Albrecht, C. Herrick, 100 Questions & Answers About Depression, Jones & Bartlett Learning, 2010.
[3] J.C. Mundt, P.J. Snyder, M.S. Cannizzaro, K. Chappie, D.S. Geralts, Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology, J. Neuroling. 20 (1) (2007) 50–64.
[4] M. Hamilton, A rating scale for depression, J. Neurol. Neurosurg. Psych. 23 (1) (1960) 56–62.
[5] N. Bogduk, Diagnostic and Statistical Manual of Mental Disorders, American Psychiatric Association, 2013.
[6] A.J. Rush, M.H. Trivedi, H.M. Ibrahim, The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression, Biol. Psych. 54 (5) (2003) 573–583.
[7] A.T. Beck, R.A. Steer, R. Ball, W.F. Ranieri, Comparison of beck depression inventories-ia and-ii in psychiatric outpatients, J. Person. Assess. 67 (3) (1996) 588–597.
[8] S.A. Montgomery, M. Asberg, A new depression scale designed to be sensitive to change, Brit. J. Psych. 134 (4) (1979) 382–389.
[9] K. Kroenke, R.L. Spitzer, The phq-9: a new depression diagnostic and severity measure, Psych. Annals 32 (9) (2002) 509–515.
[10] K. Kroenke, T.W. Strine, R.L. Spitzer, J.B. Williams, J.T. Berry, A.H. Mokdad, The phq-8 as a measure of current depression in the general population, J. Affect. Disord. 114 (1) (2009) 163–173.
[11] L.-S. Low, M. Maddage, M. Lech, L. Sheeber, N. Allen, Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents, in: 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), IEEE, 2010, pp. 5154–5157.
[12] N. Cummins, J. Epps, M. Breakspear, R. Goecke, An investigation of depressed speech detection: features and normalization, in: Interspeech, 2011, pp. 2997–3000.
[13] J.R. Williamson, T.F. Quatieri, B.S. Helfer, R. Horwitz, B. Yu, D.D. Mehta, Vocal biomarkers of depression based on motor incoordination, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM, 2013, pp. 41–48.
[14] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, Avec 2013: the continuous audio/visual emotion and depression recognition challenge, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM, 2013, pp. 3–10.
[15] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, Avec 2014: 3d dimensional affect and depression recognition challenge, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 3–10.
[16] Y. Yang, C. Fairbairn, J.F. Cohn, Detecting depression severity from vocal prosody, IEEE Trans. Affect. Comput. 4 (2) (2013) 142–150.
[17] D.R. Ladd, K.E. Silverman, F. Tolkmitt, G. Bergmann, K.R. Scherer, Evidence for the independent function of intonation contour type, voice quality, and f0 range in signaling speaker affect, J. Acoust. Soc. Am. 78 (2) (1985) 435–444.
[18] K.R. Scherer, Vocal affect expression: a review and a model for future research, Psychol. Bull. 99 (2) (1986) 143.
[19] K.R. Scherer, R. Banse, H.G. Wallbott, T. Goldbeck, Vocal cues in emotion encoding and decoding, Motiv. Emot. 15 (2) (1991) 123–148.
[20] B. Necioglu, Objectively Measurable Descriptors of Speech, Ph.D. thesis, Ph. D. dissertation, Dept. Electr. Comp. Eng., Georgia Inst. Technol., Atlanta, GA, 1998.
[21] F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in opensmile, the munich open-source multimedia feature extractor, Proceedings of the 21st ACM International Conference on Multimedia, ACM, 2013, pp. 835–838.
[22] G. Degottex, J. Kane, T. Drugman, T. Raitio, S. Scherer, Covarep-a collaborative voice analysis repository for speech technologies, 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2014, pp. 960–964.
[23] S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, H. Zen, Speech Signal Processing Toolkit (sptk), version 3.3 (2009).
[24] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., The kaldi speech recognition toolkit, in: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, no. EPFL-CONF-192584, IEEE Signal Processing Society, 2011.
[25] B. Mathieu, S. Essid, T. Fillon, J. Prado, G. Richard, Yaafe, an easy to use and efficient audio feature extraction software, in: ISMIR, 2010, pp. 441–446.
[26] F. Eyben, M. Wöllmer, B. Schuller, Openear-introducing the munich open-source emotion and affect recognition toolkit, in: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, ACII 2009, IEEE, 2009, pp. 1–6.
[27] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.
[29] M. Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1717–1724.
[30] Y. Zhang, W. Chan, N. Jaitly, Very deep convolutional networks for end-to-end speech recognition, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2017, pp. 4845–4849.
[31] L.G. Hafemann, L.S. Oliveira, P. Cavalin, Forest species recognition using deep convolutional neural networks, in: 2014 22nd International Conference on Pattern Recognition (ICPR), IEEE, 2014, pp. 1103–1107.
[32] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, Avec 2016: Depression, mood, and emotion recognition workshop and challenge, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 3–10.
[33] F. Ringeval, B. Schuller, M. Valstar, et al., AVEC 2017: Real-life depression, and affect recognition workshop and challenge, Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, ACM, 2017, pp. 3–9.
[34] I. Moore, Elliot, M.A. Clements, J.W. Peifer, L. Weisser, Critical analysis of the impact of glottal features in the classification of clinical depression in speech, IEEE Trans. Bio-Med. Eng. 55 (1) (2008) 96–107.
[35] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Commun. 71 (2015) 10–49.
[36] J.F. Cohn, T.S. Kruez, I. Matthews, Y. Yang, M.H. Nguyen, M.T. Padilla, F. Zhou, F. De, la Torre, Detecting depression from facial actions and vocal prosody, in: International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009, pp. 1–7.
[37] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, J. Epps, Diagnosis of depression by behavioural signals: a multimodal approach, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM, 2013, pp. 11–20.
[38] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, Y. Wang, Depression recognition based on dynamic facial and vocal expression features using partial least square regression, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM, 2013, pp. 21–30.
[39] E. Sánchez-Lozano, P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, J.L. Alba-Castro, Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex, Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, ACM, 2013, pp. 31–40.
[40] A. Jan, H. Meng, Y.F.A. Gaus, F. Zhang, S. Turabzadeh, Automatic depression scale prediction using facial expression dynamics and regression, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 73–80.
[41] V. Jain, J.L. Crowley, A.K. Dey, A. Lux, Depression estimation using audiovisual features and fisher vector encoding, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 87–91.
[42] M. Sidorov, W. Minker, Emotion recognition and depression diagnosis by acoustic and visual features: a multimodal approach, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 81–86.
[43] H. Perez, H.J. Escalante, L. Villasenor-Pineda, et al., Fusing affective dimensions and audio-visual features from segmented video for depression recognition, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 49–55.
[44] R. Gupta, S.S. Narayanan, Predicting affective dimensions based on self assessed depression severity, in: INTERSPEECH, 2016, pp. 1427–1431.
[45] M. Kächele, M. Schels, F. Schwenker, Inferring depression and affect from application dependent meta knowledge, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 41–48.
[46] M. Senoussaoui, M. Sarria-Paja, J.F. Santos, T.H. Falk, Model fusion for multimodal depression classification and level detection, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 57–63.
[47] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, S. Narayanan, Multimodal prediction of affective dimensions and depression in human-computer interactions, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 33–40.
[48] V. Mitra, E. Shriberg, M. McLaren, A. Kathol, C. Richey, D. Vergyri, M. Graciarena, The sri avec-2014 evaluation system, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 93–101.
[49] J.R. Williamson, T.F. Quatieri, B.S. Helfer, G. Ciccarelli, D.D. Mehta, Vocal and facial biomarkers of depression based on motor incoordination and timing, Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 65–72.
[50] L. Yang, D. Jiang, L. He, E. Pei, M.C. Oveneke, H. Sahli, Decision tree based depression classification from audio video and language information, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 89–96.
[51] X. Ma, H. Yang, Q. Chen, D. Huang, Y. Wang, Depaudionet: an efficient deep model for audio based depression classification, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 35–42.

[52] 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, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 27–34.
[53] J.R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, T.F. Quatieri, Detecting depression using vocal, facial and semantic communication cues, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 11–18.
[54] M. Nasir, A. Jati, P.G. Shivakumar, S. Nallan Chakravarthula, P. Georgiou, Multimodal and multiresolution depression detection from speech and facial landmark features, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 43–50.
[55] Z. Huang, B. Stasak, T. Dang, K. Wataraka Gamage, P. Le, V. Sethu, J. Epps, Staircase regression in OA RVM, data selection and gender dependency in AVEC 2016, Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, ACM, 2016, pp. 19–26.
[56] L. Liu, S. Lao, P.W. Fieguth, Y. Guo, X. Wang, M. Pietikäinen, Median robust extended local binary pattern for texture classification, IEEE Trans. Image Process. 25 (3) (2016) 1368–1381.
[57] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.
[58] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[59] Y. Zhu, Y. Shang, Z. Shao, G. Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics, IEEE Trans. Affect. Comput. (2017).