Multimodal Measurement of Depression Using Deep Learning Models

Le Yang
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
yangle.cst@gmail.com

Dongmei Jiang
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
jiangdm@nwpu.edu.cn

Xiaohan Xia
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
xiaohanxia@mail.nwpu.edu.cn

Ercheng Pei
NPU-VUB Joint AVSP Research lab
School of Computer Science
Northwestern Polytechnical University(NPU)
Shaanxi Key Lab on Speech and Image Information Processing
127 Youyi Xilu, Xi'an 710072 China
peiercheng@mail.nwpu.edu.cn

Meshia Cédric Oveneke
NPU-VUB Joint AVSP Research lab
Dept. Electronics & Informatics (ETRO)
Vrije Universiteit Brussel(VUB)
Pleinlaan 2, 1050 Brussels, Belgium
mcovenek@etro.vub.ac.be

Hichem Sahli
NPU-VUB Joint AVSP Research lab
Deptartment of Electronics & Informatics(ETRO)
Vrije Universiteit Brussel(VUB)
Pleinlaan 2, 1050 Brussels, Belgium
Interuniversity Microelectronics Centre(IMEC)
Kepeldreef 75, 3001 Heverlee, Belgium
hsahli@etrovub.be

ABSTRACT

This paper addresses multi-modal depression analysis. We propose a multi-modal fusion framework composed of deep convolutional neural network (DCNN) and deep neural network (DNN) models. Our framework considers audio, video and text streams. For each modality, handcrafted feature descriptors are input into a DCNN to learn high-level global features with compact dynamic information, then the learned features are fed to a DNN to predict the PHQ-8 scores. For multi-modal fusion, the estimated PHQ-8 scores from the three modalities are integrated in a DNN to obtain the final PHQ-8 score. Moreover, in this work, we propose new feature descriptors for text and video. For the text descriptors, we select the participant's answers to the questions associated with psychoanalytic aspects of depression, such as sleep disorder, and make use of the Paragraph Vector (PV) to learn the distributed representations of these sentences. For the video descriptors, we propose a new global descriptor, the Histogram of Displacement Range (HDR), calculated directly from the facial landmarks to measure their displacements and speed. Experiments have been carried out on the AVEC2017 depression sub-challenge dataset. The obtained results show that the proposed depression recognition framework obtains very promising accuracy, with the root mean square error (RMSE) as 4.653, mean absolute error (MAE) as 3.980 on the development set, and RMSE as 5.974, MAE as 5.163 on the test set.

CCS CONCEPTS

• **Pattern Recognition → Applications|signal processing, computer vision, speech processing**;

KEYWORDS

Depression recognition, DCNN-DNN, multi-modal

ACM Reference format:
Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2017. Multimodal Measurement of Depression Using Deep Learning Models. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge, Mountain View, CA, USA, October 23, 2017 (AVEC'17),* 7 pages.
https://doi.org/10.1145/3133944.3133948

AVEC'17, October 23, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-5502-5/17/10…$15.00
https://doi.org/10.1145/3133944.3133948

1 INTRODUCTION

Depression is a state of low mood and aversion to activity that can affect a person's thoughts, behaviours, feelings, and sense of well-being. At present, depression and anxiety disorders are highly prevalent worldwide causing burden and disability for individuals, families and society [7].

Accurately estimating and evaluating the level of depression has very broad application prospects. Within the past several years, various models have been investigated for depression level estimation from behavioral observations [2], [3], [12]. The most commonly used models, such as Gaussian Mixture Models (GMM), Support Vector Machines (SVM) and Relevance Vector Regression (RVR), have achieved remarkable performance. In [1], Alghowinem *et al.* compared four popular classifiers, namely GMM, SVM, Multilayer Perceptron neural networks (MLP), and Hierarchical Fuzzy Signature (HFS), for the task of depression detection. In [18], the authors modelled a wide range of vocal features using a variety of approaches, including support vector regression (SVR), GMM and decision trees. A staircase Gaussian approach, in which each GMM is comprised of an ensemble of Gaussian classifiers, has been proposed in [26] and [25]. The model outperformed most of the approaches used for the Audio Visual Emotion Challenge (AVEC) 2014 and AVEC 2016.

These studies, however, used relatively shallow architectures. To increases the recognition robustness, deep learning has emerged as alternative to such methods. Deep Convolutional Neural Network (DCNN) has been increasingly used and has demonstrated impressive performances on different tasks such as object recognition [24], object detection [8], face verification [21], emotion recognition [17] [6] [27] and depression classification [16] [29]. It can not only characterize large data variations but also learn compact and discriminative feature representation, especially when the size of data is large. On the other hand, deep neural networks (DNNs) have also been widely adopted for several recognition tasks as well as for discrete and continuous affect recognition [15] [13] [11]. Experimental results demonstrated that due to their strong ability of capturing the underlying nonlinear relationship among data with multiple layers, DNN improves the recognition or classification accuracy greatly.

Inspired by the powerful learning ability of deep models, we present a multi-modal hybrid deep learning framework composed of DCNN and DNN to learn, from each modality, low dimensional global feature vectors with compact dynamic information and improve the prediction accuracy of the PHQ-8 scores. For each unimodal (audio, video, text), the DCNN-DNN is comprised of the following two steps. First, features descriptors extracted from the considered modality are input to a DCNN to learn high-level features. The learned features are then fed to a DNN, which is trained to predict the PHQ-8 scores. For the final fusion, the estimated PHQ-8 scores from the three modalities are integrated in a fusion network built with a DNN, which is also trained to predict the final PHQ-8 scores.

Additionally, we propose new features descriptors for text and video. Analyzing text information has emerged as a new approach in studying emotion-related textual indicators as well as depression symptoms [19] [20] [28]. In our framework, we consider the interaction transcripts, from which we select the patient's answers to the questions related to psychoanalytic symptoms of depression, such as sleeping disorder, feelings, etc.. To represent the selected sentences, we make use of the Paragraph Vector (PV) descriptor, which has been introduced in [14] to directly learn the distributed representations of sentences and documents. Different from most of the conventional text feature extraction, such as Bag-of-words, PV takes consideration of context semantics using low-dimensional representations. For the video descriptors, we exploit the detected 2D landmarks of the face and propose a new global descriptor denoted as Histogram of Displacement Range (HDR). To construct a HDR, each landmark displacement votes, with its lengths in the horizontal and vertical directions, in a displacement range histogram.

In this paper, we target the Depression Sub-Challenge (DSC) task of AVEC2017 [5]. Experiments are carried out on the Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) database [9], results indicate that our hybrid DCNN-DNN framework presents promising performance.

The outline of the paper is as follows. The feature descriptors used as inputs to the DCNNs are introduced in Section 2. Section 3 gives an overview of the proposed DCNN-DNN based multi-modal depression prediction framework. Section 4 illustrates and analyzes the experimental results. Finally, conclusions and future works are given in Section 5.

2 FEATURE DESCRIPTORS

The collected AVEC2017 depression dataset include audio, video and transcripts (text) of the conversations between the participants and an animated virtual interviewer. Each conversation session is composed of several segments, each corresponding to a question-answer pair. In the following sections, we describe the feature sets we use from each modality.

2.1 Text Descriptors

The transcript files provided by the DAIC-WOZ include time stamp and dialogue contents. An example is given in Table 1. We conducted content analysis of the transcripts to select the patient's answers to the questions related to psychoanalytic aspects associated to depression symptoms. Five aspects are considered in this study:

(1) Prior depression diagnoses
(2) Prior post-traumatic stress disorder (PTSD) diagnosis
(3) Sleep disorder
(4) Feeling
(5) Personality.

To further use the selected sentences for PHQ-8 prediction, we make use of the recently proposed Paragraph Vector (PV) embedding model [14]. To the best of our knowledge, our approach is the first which attempt to apply PV to textual transcripts for depression analysis.

The Paragraph Vector [14] model is an extension of the Word2Vec model of distributed representations of words in a vector space. The Paragraph Vector can generate the representation of any sentence without considering the length of text. The objective of both models is to maximize the average log probability of any given word in a sequence of training words $w_1, w_2, w_3, \cdots, w_n$, conditioned on the appearance of other words of the same sequence [14].

$$\frac{1}{n}\sum_{t=k}^{n-k}\log p(w_t|w_{t-k},...,w_{t+k})$$

Table 1: Example of the question-answer pairs in transcript files

start_time	stop_time	speaker	value
...	...	...	...
148.94	150.8	Ellie	do you consider yourself an introvert?
153.04	155.01	Participant	um i was an extrovert
...	...	...	...
326.63	328.83	Ellie	how easy is it for you to get a good night's sleep?
329.53	331.41	Participant	mm it isn't
...	...	...	...
384.442	386.302	Ellie	have you been diagnosed with depression?
386.952	387.372	Participant	yes
...	...	...	...
438.69	440.071	Ellie	how have you been feeling lately?
442.61	443.52	Participant	i guess sorta depressed
...	...	...	...

The above definition describes a prediction task, which is usually solved via the usage of a multiclass classifier such as *softmax*.

$$p(w_t|w_{t-k},...,w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \quad (1)$$

where each term y_i is the un-normalized log-probability for each output word i, computed as:

$$y = b + Uh(w_{t-k},...,w_{t+k};W) \quad (2)$$

where U, b are the softmax parameters and the function h is constructed by a concatenation (or average) of the word vectors extracted from the matrix W.

Figure 1 illustrates the Paragraph Vector model. In the PV model, in addition to mapping each vocabulary word to a column in W, every paragraph is subsequently mapped to a unique vector, represented by a column in a new document matrix D. The paragraph vector and word vectors are averaged and subsequently fed to the related system for the prediction of next word W_{i+1}. The paragraph token acts as a memory that remembers what is missing from the current context, i.e. the topic of the paragraph. The algorithm operates in two stages: unsupervised training to get word vectors W and the inference stage to get paragraph vectors D, which are used as descriptors of the sentences.

An important advantage of Paragraph Vectors is that they are learned from unlabeled data. They also inherit an important property of the Word2vec model, being the semantics of the words. Moreover, they take into consideration the word order.

2.2 Video Descriptors

In this paper, we exploit the 2D landmarks of the face provided by AVEC2017 and propose a new global descriptor denoted as Histogram of Displacement Range (HDR). To construct HDR, each landmark displacement votes with its length, in the horizontal and vertical directions, in a displacement range histogram. The landmark displacements are described by a histogram of n equally spaced bins, $R_i, i = 1, ..., n$, spanning the range $[-30, 30]$ pixels. For

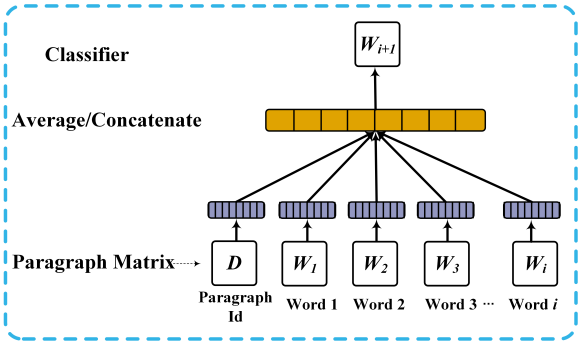

Figure 1: A framework for learning paragraph vector (adapted from [14]). The concatenation or average of the W_1, W_2, W_3, W_i and D are used to predict the W_{i+1}. D is called as paragraph matrix, in which each column is a paragraph vector. The paragraph vector can represents the topic of this paragraph.

temporal modeling, we estimate the landmarks displacements in the horizontal and vertical directions, at different time intervals $M_k, k = 1, \cdots, K$ and concatenate the obtained histograms. Formally, let $D_x(i,j) = j_{i+M_k}(x) - j_i(x)$ and $D_y(i,j) = j_{i+M_k}(y) - j_i(y)$ be the horizontal and vertical displacement of landmark $j(x,y)$ between frame i and frame $i + M_k$. Each bin of the histogram contains the number of occurrences of displacements in the corresponding range. HDR records the range and speed of the displacements of the facial landmarks in the horizontal and vertical directions. Algorithm 1 summarizes the HDR estimation. For our experiments we

Algorithm 1: Histogram of Displacement Range

input:
Set time interval $M := \{M_1, M_2, M_3, ..., M_k\}$;
Set the range $R := \{R_1, R_2, R_3, ..., R_n\}$;
Time series features: Landmarks;
output:
Histogram of Displacement Range: $N_{(M_k,j,R_n)}$;

```
for each M_k do
  for each landmark j(x, y) and
    i ≤ totalframe − M_k do
    Dx(i, j) := j_{i+M_k}(x) - j_i(x);
    Dy(i, j) := j_{i+M_k}(y) - j_i(y);
    for each R_n do
      Num_x := Dx(i, j) ∈ R_n;
      Num_y := Dy(i, j) ∈ R_n;
      Num_x := Num_x/(totalframe − M_k);
      Num_y := Num_y/(totalframe − M_k);
      N_(M_k,j,R_n) := [N_(M_k,j,R_n) Num_x Num_y];
    end
  end
end
```

considered 5 time intervals (M_k), as 10, 20, 30, 40, and 50 frames, respectively. In our experiment, we find that males show less head

movements and facial landmarks movements than females, therefore, with respect to the histogram bins, we considered 6 equally spaced bins (R_i), spanning the range $[-30, 30]$ pixels for females, and 4 equally spaced bins spanning the range $[-20, 20]$ pixels for males. In total we obtained 4080 HDR features for females, and 2720 for males.

2.3 Audio Descriptors

In this work, considering the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [4] and the INTERSPEECH Challenges [22] feature sets, we utilize the openSMILE toolkit [23] to extract for each audio segment, 238 low level descriptors (LLDs), comprising 211 spectral and energy related features and 27 voicing related dynamic features. The LLDs are shown in Table 2 where the numbers between brackets are the dimensions of the extracted features vectors, and $\triangle$, $\triangle\triangle$ denote the first and second order derivatives, respectively. The LLDs are extracted with the frame length of 60ms and frame shift of 10ms. 25 statistical functionals and 4 regression functionals, as shown in Table 3, have been performed on the extracted LLDs, resulting in a 6902 dimensional feature vector for each speech segment. These features correspond to the mostly used features for speech emotion recognition.

Table 2: Low level descriptors from openSMILE (238)

Energy and Spectral related (211)	
PLPCC(5)+$\triangle$+$\triangle\triangle$	MFCC-sma(15)+$\triangle$+$\triangle\triangle$
LOGenergy(1)+$\triangle$+$\triangle\triangle$	Chroma(12)
LspFreq(8)+$\triangle$	SpectralRollOff(4)+$\triangle$
LengthL1norm(2)+$\triangle$	LpcCoeff(11)
LogRelF0(2)+$\triangle$	Amplitude(3)+$\triangle$
SpectralEnergy(2)+$\triangle$	SpectralSlope(2)+$\triangle$
Zcr(1)+$\triangle$	Loudness(1)+$\triangle$
RASTA-filtered(26)+$\triangle$	RMSenergy(1)+$\triangle$
Spectral(Flux, Centroid, Entropy,	Hammarberg(1)+$\triangle$
Variance, Skewness, Kurtosis,	LpGain(1)
Harmonicity, Flatness)(8)+$\triangle$	AlphaRatio(1)+$\triangle$
Voicing related (27)	
F0(2)+$\triangle$ LogHNR(1)+$\triangle$	JitterLocal(1)+$\triangle$
FormantFreqLpc(6)	ShimmerLocal(1)+$\triangle$
FormantBandwidthLpc(6)	FormantFrameIntensity(1)
VoicingFinalUnclipped(1)+$\triangle$	JitterDDP(1)+$\triangle$

Table 3: Functionals from openSMILE (29)

Statistical functionals (25)
max, min, arithmetic mean, norm, variance, stddev, skewness, kurtosis, numPeaks, meanPeakDist, peakMean, peakMeanMeanDist, quartiles(1-3), samplepos, numSegments, meanSegLen, maxSegLen, minSegLen, upleveltime25, upleveltime50, upleveltime75, risetime, falltime
Regression functionals (4)
linregc1, linregc2, linregerrA, linregerrQ

3 MULTI-MODAL DEPRESSION RECOGNITION FRAMEWORK

The audio, visual, and the text networks are first trained individually using the ground truth labels. We first train the DCNN model. For each audio, video and text segments we pass the corresponding feature descriptors through a DCNN, as shown in Figure 2. The DCNN has n convolutional layers, followed by one ReLU, Pooling and Dropout layers, while the last convolutional layer is followed by two fully connected layers. For the training we add a fully-connected layer to produce the prediction score. The loss function associated to the output of the model is the Euclidean loss. After training of DCNN, we freeze the weights, discard the last layer and connect the second fully connected layer to the visible layer of the DNN. Here also, the loss function associated to the output of the model is the Euclidean loss. In our current implementation, the two networks are trained separately, without back-propagating the DNN loss to the DCNN.

As described in Figure 3, our hybrid deep learning model contains three input streams: the audio/video network processing audio/video data with two DCNN-DNN models, and the text network processing text data with 5 DCNN-DNN models connected to a fusion DNN model. The outputs of these three networks are fused in a fusion network built with a DNN model. The structural hyper-parameters of the used DCNNs and DNNs are given in Section 4.3. All of our DCNN and DNN models were trained using Caffe [10].

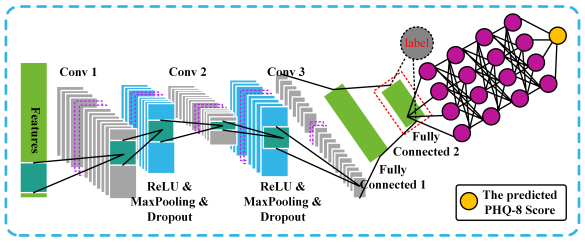

Figure 2: Unimodal DCNN-DNN model for depression recognition.

4 EXPERIMENTS AND ANALYSIS

In this section, we start by giving a brief description of the used dataset, followed by the analysis of experimental results and comparison with the state-of-the-art methods.

4.1 The AVEC 2017 Depression Dataset

The AVEC2017 depression dataset [9] consists of 189 segments of clinical interviews designed to support the diagnosis of psychological distress conditions such as anxiety, depression, and post-traumatic stress disorder. A segment corresponds to an audio/video recording of the participant's answering a question from Ellie, the animated virtual interviewer. The recorded segments were split into a training set of 107 segments, a development set of 35 segments and a test set composed of 47 segments. For each segment of the training and development sets, AVEC2017 provides the PHQ-8 level of depression and a binary value of depressed/not-depressed.

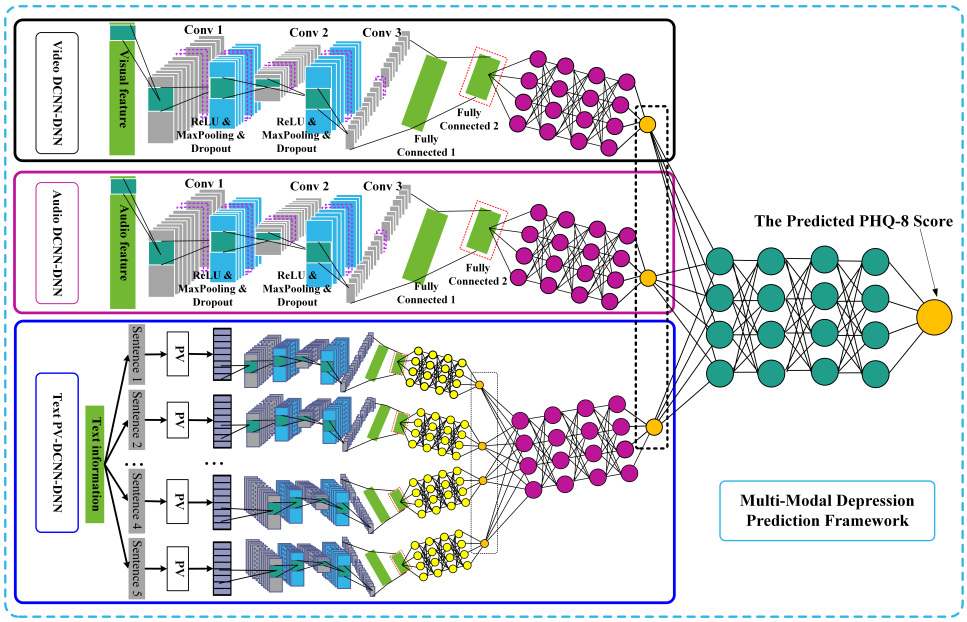

Figure 3: The structure of our proposed hybrid deep model for audio-visual-text depression recognition

Table 4: Data distribution of the depressed/not depressed on the training set(in the brackets are after sampling).

Gender	Depressed Sample	Not Depressed Sample
Female	17(48)	27(46)
Male	13(40)	50(48)

4.2 Data Balancing

The training set of the AVEC2017 consists of imbalanced *depressed* and *not-depressed* samples, as shown in Table 4. Such imbalanced data may decrease the recognition performance and cause overfitting. In this work we re-sample the dataset to obtain a more balanced data. The sampling is made as follows: we first remove the *non-speaking* segments when the participant listens to Ellie. Then for the *speaking* segments from *depressed* samples, the longest four segments in each session are taken as four new samples for female (five for male), and for *not-depressed* samples, the longest four segments of each PHQ-8 level are taken as four new samples for both female and male. Finally, for the females we obtain 46 segments as *not-depressed* samples and 48 segments of *depressed* samples. For the males, we obtain 48 *not-depressed* samples and 40 *depressed* samples. The distribution of the sampled data set is shown in the brackets of Table 4.

4.3 Model Parameters

The Paragraph Vector model can be trained from a large text corpus. In this work, we train the Paragraph Vector model with 402, 325 dialogues collected from *Friends*, *The Big Bang Theory* and *Game of Thrones*. Most of these dialogues are simple and short. The parameters used in Paragraph Vector model are the dimension L of the PV vector for each sentence, and the window size S. In our experiments we evaluated $L \in \{50, 100, 150\}$ and $S \in \{5, 10\}$.

We trained gender specific hybrid deep models. For designing the DCNN-DNN networks, we have tested several network architectures and selected the ones which provided the best PHQ-8 prediction, in terms of root mean square error (RMSE) and mean absolute error (MAE), on the development set. For the DCNNs, we evaluated architectures with 3, 4 and 5 convolutional layers (CV). The convolution kernel size has been set to $1x5$ and $stride = 1$, and

Table 5: The best model structure and parameters.

Audio/Visual Single Modal Depression Recognition				
Gender	Modal	DCNN		DNN
		Conv	Fully-Conn	Hidden Layers
Female	audio	{24,48,64}	{100,30}	{100,100,100,30}
Female	video	{24,48,64}	{100,30}	{100,100,100,30}
Male	audio	{24,48,64}	{50,20}	{150,150,150,50}
Male	video	{12,48,64}	{100,20}	{150,150,150,50}

Text Based Depression Recognition (For all sentences, we adopt the same parameters.)					
PV		DCNN		Single-stream DNN	Multi-stream DNN
L	S	Conv	Fully-Conn	Hidden Layers	Hidden Layers
100	5	{64,48,48}	{64,24}	{100,100,30}	{35,35,35,35}

Audio-Video-Text Multi-modal Depression Recognition	
Gender	The Final Fusion DNN
Female	{35,35,35,35,15}
Male	{30,30,30,30,15}

the number of feature maps $N_{maps} \in \{12, 24, 48, 64, 128\}$. Each CV layer is followed by a ReLU, MaxPooling and Dropout layer, except for the last CV. The pooling kernel size was set to $2x2$, $stride = 2$, and for the Dropout layer, we evaluated dropout ratios in the range $[0.3, 0.5]$.

Once the DCNN model is trained, the outputs of the second fully-connected layer are fed into the DNN model, whose last layer is still the PHQ-8 score. We evaluated several DNN layers, $N_{dlayer} \in \{3, 4, 5\}$ and for each layer different number of hidden nodes have been tested, $N_{nodes} \in \{15, 20, 25, 30, 35, 50, 100, 150\}$. The ReLU is used as activation function, the Euclidean loss as loss function.

In the multi-modal depression estimation framework, the estimated PHQ-8 scores from single modalities are input together into another DNN model in which the number of layers $N_{dlayer} \in \{2, 3, 4\}$, and in each layer, the number of hidden nodes is set as $N_{nodes} \in \{2, 4, 6, 8, 10, 15, 20, 25, 30, 35\}$.

The experiments are implemented on a Tesla C2075 GPU machine. Because of the very limited size of the dataset, the time consumed for training is very low. We selected the architectures which provided the best accuracy, as listed in Table 5, and the nodes of each hidden layer are shown in the braces.

4.4 Depression Recognition Results

4.4.1 Single-modal depression recognition. Single modal prediction results of the PHQ-8 scores for females and males on the development set are given in Table 6. For the audio and video we compared the proposed DCNN-DNN to a DCNN model for PHQ-8 prediction. From Table 6, one can notice that: 1) The DCNN-DNN framework, outperforms the (Audio/Video)-DCNN. The same conclusion also applies to text, although we do not give the results of DCNN in the table. 2) For the text modality, Sentence-1 produces the best PHQ-8 results. This is reasonable because Sentence-1 relates to a previous depression diagnosis, and the answer is normally very straight forward as "yes" or "no". While for other sentences, such as Sentence 3 (Sleep disorder), the answers are quite complex or ambiguous, which make the PV features not so discriminative. This explains why the fusion scheme of text does not significantly improve the performance over Sentence 1.

4.4.2 Multi-modal depression recognition. Multi-modal PHQ-8 score estimations for females, males and all participants (put male and female results together) are given in Table 7. For both development and test sets our proposed DCNN-DNN fusion framework in Figure 3 obtains better performance than the baseline. We therefore believe that our approach can be a good solution to predict PHQ-8 scores from audio, video and text data.

Table 6: Single-modal recognition of PHQ-8 for female and male on the development set

Modal	Feature-model	Female		Male	
		RMSE	MAE	RMSE	MAE
Audio	Audio-DCNN	5.744	**4.484**	5.736	5.241
	Audio -DCNN-DNN	**5.669**	4.597	**5.590**	**5.107**
Video	Video(HDR) -DCNN	5.325	**4.299**	5.879	5.500
	Video(HDR) -DCNN-DNN	**5.199**	4.329	**5.606**	**5.174**
Text	Sentence 1 -DCNN-DNN	**4.361**	3.750	**4.406**	**3.525**
	Sentence 2 -DCNN-DNN	5.168	3.931	6.112	5.283
	Sentence 3 -DCNN-DNN	6.349	5.238	6.627	6.118
	Sentence 4 -DCNN-DNN	5.789	5.139	6.644	6.165
	Sentence 5 -DCNN-DNN	6.093	4.937	4.876	3.875
	DNN Fusion	4.381	**3.307**	4.659	3.741

Table 7: Depression Recognition Results

Gender	Data set	RMSE	MAE
Female	training	2.327	1.823
	Dev.	4.702	3.873
Male	training	3.047	2.673
	Dev.	4.594	4.107
All	Dev.(baseline)	6.620	5.520
	Dev.(proposed)	**4.653**	**3.980**
	test.(baseline)	6.970	6.120
	test.(proposed)	**5.974**	**5.163**

5 CONCLUSIONS

This paper provides an overview of our proposed audio video and text fusion framework designed for AVEC2017 Depression Challenge, as well as the experimental results that outperform the baseline results on the test and development sets. Besides the DCNN and DNN based fusion framework, we propose a new video descriptor HDR based on the tracked 2D facial landmarks, and a Paragraph Vector embedding as text descriptor. Experimental results on the AVEC2017 depression dataset show that our proposed multimodal hybrid framework integrating DCNN and DNN models obtains promising performance.

In our future work, we will include an improved text analysis model to account for all the interview conversations. Indeed, for the single modality, text alone with sentence-1 produces the best results, hence more analysis of the fusion scheme will be made in future research. . End-to-end learning strategies will also be investigated to further boost the depression recognition performance.

ACKNOWLEDGMENTS

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), and the VUB Interdisciplinary Research Program through the EMO-App project.

REFERENCES

[1] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Tom Gedeon, Michael Breakspear, and Gordon Parker. 2013. A comparative study of different classifiers for detecting depression from spontaneous speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 8022–8026.

[2] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 1–7.

[3] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke. 2011. An investigation of depressed speech detection: Features and normalization. In *Twelfth Annual Conference of the International Speech Communication Association*.

[4] Florian Eyben, Klaus R. Scherer, BjÃűrn W. Schuller, Johan Sundberg, Elisabeth AndrÃĬ, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, and Shrikanth S. Narayanan. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing* 7, 2 (2016), 190–202.

[5] Ringeval Fabien, Bjorn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Mozgai Sharon, Cummins Nicholas, Schmitt Maximilian, , and Maja Pantic. 2017. AVEC 2017 - Real-life Depression, and Affect Recognition Workshop and Challenge. In *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge*.

[6] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 445–450.

[7] Lynne Friedli, World Health Organization, et al. 2009. Mental health, resilience and inequalities. (2009).

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[9] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The Distress Analysis Interview Corpus of human and computer interviews.. In *LREC*. 3123–3128.

[10] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093* (2014).

[11] Yelin Kim, Honglak Lee, and Emily Mower Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 3687–3691.

[12] Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology* 15, 2 (2012), 99–117.

[13] Duc Le and Emily Mower Provost. 2013. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 216–221.

[14] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[15] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. 2013. Hybrid Deep Neural Network–Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 312–317.

[16] Xingchen Ma, Hongyu Yang, Qiang Chen, Di Huang, and Yunhong Wang. 2016. DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 35–42.

[17] Veena Mayya, Radhika M Pai, and MM Manohara Pai. 2016. Automatic Facial Expression Recognition Using DCNN. *Procedia Computer Science* 93 (2016), 453–461.

[18] Vikramjit Mitra, Elizabeth Shriberg, Mitchell McLaren, Andreas Kathol, Colleen Richey, Dimitra Vergyri, and Martin Graciarena. 2014. The SRI AVEC-2014 evaluation system. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 93–101.

[19] Michelle Renee Morales and Rivka Levitan. 2016. Speech vs. text: A comparative analysis of features for depression detection systems. In *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 136–143.

[20] Anastasia Pampouchidou, Olympia Simantiraki, Amir Fazlollahi, Matthew Pediaditis, Dimitris Manousos, Alexandros Roniotis, Georgios Giannakakis, Fabrice Meriaudeau, Panagiotis Simos, Kostas Marias, et al. 2016. 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, 27–34.

[21] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

[22] BjÃűrn Schuller, Stefan Steidl, Anton Batliner, Elmar NÃűth, Alessandro Vinciarelli, Felix Burkhardt, Rob Son, Felix Weninger, Florian Eyben, and Tobias Bocklet. 2012. The INTERSPEECH 2012 speaker trait challenge. In *INTERSPEECH 2012, Conference of the International Speech Communication Association*.

[23] Björn W Schuller, Stefan Steidl, Anton Batliner, Julien Epps, Florian Eyben, Fabien Ringeval, Erik Marchi, and Yue Zhang. 2014. The INTERSPEECH 2014 computational paralinguistics challenge: cognitive & physical load.. In *INTERSPEECH*. 427–431.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[25] James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. 2016. Detecting Depression using Vocal, Facial and Semantic Communication Cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 11–18.

[26] James R Williamson, Thomas F Quatieri, Brian S Helfer, Gregory Ciccarelli, and Daryush D Mehta. 2014. 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, 65–72.

[27] Jingwei Yan, Wenming Zheng, Zhen Cui, Chuangao Tang, Tong Zhang, Yuan Zong, and Ning Sun. 2016. Multi-clue fusion for emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 458–463.

[28] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. 2016. Decision Tree Based Depression Classification from Audio Video and Language Information. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 89–96.

[29] Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo. 1949. Automated Depression Diagnosis based on Deep Networks to Encode Facial Appearance and Dynamics. *IEEE Transactions on Affective Computing* PP, 99 (1949), 1–1.