Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics

Yu Zhu, Yuanyuan Shang, Zhuhong Shao, and Guodong Guo, *Senior Member, IEEE*

Abstract—As a severe psychiatric disorder disease, depression is a state of low mood and aversion to activity, which prevents a person from functioning normally in both work and daily lives. The study on automated mental health assessment has been given increasing attentions in recent years. In this paper, we study the problem of automatic diagnosis of depression. A new approach to predict the Beck Depression Inventory II (BDI-II) values from video data is proposed based on the deep networks. The proposed framework is designed in a two stream manner, aiming at capturing both the facial appearance and dynamics. Further, we employ joint tuning layers that can implicitly integrate the appearance and dynamic information. Experiments are conducted on two depression databases, AVEC2013 and AVEC2014. The experimental results show that our proposed approach significantly improve the depression prediction performance, compared to other visual-based approaches.

Index Terms—Automated depression diagnosis, nonverbal behavior, deep convolutional neural networks, flow dynamics

1 Introduction

Major depression disorder (MDD) is one of the prevalent causes of disability which heavily threatens the mental health of human among all age groups [1]. Depression disorder, with 10-20 percent for women and 5-12 percent for men lifetime risk, can severely affect person's thoughts, behavior, feelings, and ability to work. Depressed people may feel sad, helpless, anxious, hopeless, worried, irritable, or restless, even in the worst scenario, severe depression could even lead to suicide [2], [3]. Fortunately, through proper medication, psychological counseling and other clinical methods, MDD is treatable despite of its severity. Currently, the diagnosis of MDD mostly requires comprehensive assessment by experienced professional. It is largely constrained by individual subjective observation and lack of real-time measurements. As the increasing number of people suffering from MDD, it also brings the burden to accurate diagnosis. Therefore, machine learning based methods are expected to provide an objective assessment and fast diagnosis, which can aid the MDD therapy.

The study on automated mental health assessment has been given increasing attention in recent years. One way of keeping track of patients with depression is online monitoring through human computer interaction and affective computing technologies. Particularly, machine learning methods can be used to analyze affect and expressive behaviours that are directly related to depression. Evidence has shown that speech production differs in people with depression [4], [5], thus some methods have been proposed utilizing audio cues for depression diagnosis [6], [7], [8], [9], [10], [11]. It is also suggested that nonverbal cues are indicative of depression severity, such as gestures and expressions [12], [13]. Studies have shown that more than a half visual-based nonverbal behaviors are around the facial region in human communication activities [14], [15], [16], [17]. Accordingly, in this work we focus on the visual-based nonverbal behaviors for depression diagnosis.

From the machine learning perspective, depression diagnosis can be modeled as a regression problem, e.g., in AVEC2013 and AVEC2014 depression recognition challenges, the goal is to predict the depression score called Beck Depression Inventory-II (BDI-II score [18], see Table 1) for the subject in each video. To deal with this problem, facial appearance and dynamics in video clips are often considered very useful for depression diagnosis [19]. We study depression recognition and propose a new approach to model the facial appearance and dynamics, based on deep convolutional neutral networks (DCNN). Our approach is designed in a two-stream manner, combined with joint-tuning layers for depression prediction. Specifically, facial appearance representation is modeled through a very deep neural network, using face frames as the input. Facial dynamics are modeled by another deep neural network, with face "flow images" as the input. Face "flow images" are generated by computing within the video sub-volumes using the optical flow, to capture facial motions. The two deep networks are then integrated by joint-tuning layers into one deep network, which can further improve the overall performance. To the best of our knowledge, our proposed approach is for the first time to employ deep learning technology for depression diagnosis. Extensive experiments conducted on two depression databases, AVEC2013 [20] and AVEC2014 [21] show that, our approach achieved better results than the state-of-the-art visual-based methods for depression recognition.

The rest of the paper is organized as follows: Section 2 is about previous works on depression diagnosis. In Section 3, our proposed method and network architecture are presented in details. Next, experiments are conducted on two databases and the results are shown in Section 4. Finally, some discussions and future works are given.

2 Related Work

The Audio-Visual Emotion Challenge and Workshop 2013 and 2014 (AVEC2013 and AVEC2014) held the competition events for depression recognition as one of its sub challenges. The depression values are tested on the collected audio-video databases (see Sections 4.1 and 4.2 for more details about the databases). We focus on video-based approach, where the audio cue is not utilized. In the following, we briefly describe the competing visual based methods in the AVEC2013 and AVEC2014 competitions.

Baseline features for AVEC2013 is the Local Phase Quantization (LPQ) [22], which has shown good performance in facial expression recognition. For AVEC2013 depression recognition, face detection, fitting and alignment were performed for each video frame. Then the dense LPQ features were extracted from those facial regions. Facial feature for each frame is represented by concatenating histograms of different blocks within the face region. Finally, the Support Vector Regression (SVR) is applied for learning and prediction.

In Cummins et al. 's work [8], two different features are compared: the Space-Time Interest Points (STIPs) [23], and Pyramid of Histogram of Gradients (PHOG) [24]. In their method, face tracking is applied for each video frames to obtain the face region. Then, both STIPs and PHOG features are extracted from the aligned face

- *Y. Zhu is with Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506. E-mail: xperzy@gmail.com.*
- *G. Guo is with Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, and with Beijing Advanced Innovation Center for Imaging Technology, Beijing 100048, China. E-mail: Guodong.Guo@mail.wvu.edu.*
- *Y. Shang and Z. Shao are with the College of Information Engineering, Capital Normal University, Beijing 100048, China, and Beijing Advanced Innovation Center for Imaging Technology, Beijing 100048, China. E-mail: syy@bao.ac.cn, shaozh2015@163.com.*

Manuscript received 30 Apr. 2016; revised 7 Dec. 2016; accepted 2 Jan. 2017. Date of publication 9 Jan. 2017; date of current version 5 Dec. 2018.
Recommended for acceptance by M. Soleymani.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2017.2650899

TABLE 1
Beck Depression Inventory-II (BDI-II) Score and Depression Severity

BDI-II Score	Depression Severity
0 - 13	None
14 - 19	Mild
20 - 28	Moderate
29 - 63	Severe

images. Those features are further generated as histograms by using the bag-of-word scheme. Finally, SVR with histogram intersection kernel was used for training and testing. In their experiments, PHOG has shown better results than STIPs.

Meng et al. [7] utilized Motion History Histogram (MHH) [25] to characterize motion information of each pixel in the video. Totally there were 5 MHH based images generated from each video frame. Then Edge Orientation Histogram (EOH) and Local Binary Patterns (LBP) [26] features were extracted from each MHH based image. Finally, a Partial Least Squares (PLS) [27] regressor was used for learning a regression function. In their method, the MHH based descriptor can reflect all behaviors to some extent but the temporal information is still not well-encoded.

In our previous work [19], the temporal dynamics is captured by the LPQ-TOP features from facial region sub-volumes. Then a behavior pattern dictionary is learned through sparse coding schemes. The sparse codes are calculated for each LPQ-TOP feature separately. Finally, a discriminative mapping method and decision level fusion were applied to further improve the accuracy for depression diagnosis.

In the AVEC2014, local dynamic appearance descriptor LGBP-TOP [28] has been adopted as the baseline video features. LGBP-TOP utilizes a number of Gabor filters on a block of consecutive frames as input, then apply LBP feature extraction from three different orthogonal slice of the block: XY, XT and YT. The resulting patterns are further histogrammed and concatenated into the final feature representation. Support Vector Regression (SVR) is used for the prediction, which is the same as AVEC2013.

In the video based approach from [29], the authors detected the face within each video frame, then utilized three motion related features: motion history image, motion static image and motion average image from the detection face region. The features were also combined with the relative differences of the face and eye coordinates. Finally, the extracted features were fed into a SVR for prediction.

In [30], the authors detected and cropped the faces within each video frame, then three features were extracted: Local Binary Patterns (LBP), Edge Orientation Histogram (EOH) and Local Phase Quantization (LPQ). Instead of generating from an image sequence, they proposed an 1-D Motion History Image (MHH) that extracts the changes on each component in a feature vector sequence. Then histogram features are used to represent all the components of the feature vector in one video. Partial Least Squares (PLS) regression is applied for the final prediction.

In [31], the authors proposed to utilize both LGBP-TOP and LPQ features for video representation. They focused on the inner facial regions that correspond to eyes and mouth for feature extraction. Then the Canonical Correlation Analysis (CCA) is applied on the feature vectors, and the two features are ensembled to generate the final regression results.

Most of the above mentioned methods were based on hand-crafted features which were proposed for facial analysis or expression recognition. We want to explore a more robust representation for depression analysis, which can better capture the appearance and dynamics cues. Specifically, we investigate a new approach based on the deep learning networks, for automated depression diagnosis.

3 Deep Neural Network Architectures for Depression Recognition

Video data can be naturally viewed as two components, i.e., spatial and temporal components. For the problem of depression recognition, the spatial part carries the appearance information about the face and static expressions of the subject. On the other hand, the temporal part captures the motion across frames, containing the facial dynamics such as the expression and micro expression changes of the subject. Therefore, we explore the architecture for depression analysis, encoding both appearance and dynamics. As shown in Fig. 2, each part is implemented using a DCNN. Moreover, joint layers are proposed to combine the two streams for the final depression recognition.

3.1 Appearance-DCNN

Deep convolutional neutral networks (DCNN) are known to be very effective for learning face representations given a large number of face samples. However, for the specific task of depression recognition, usually the size of data available is very limited. To handle this issue, we utilize a cascaded way to train the facial appearance deep model in two steps: a pre-training step and a fine-tuning step.

In pre-training, a deep network (e.g., GoogLeNet [32]) is trained from scratch by utilizing a public face recognition database [33] which contains 494,414 facial images from 10,575 subjects. After this step, the deep network is expected to effectively capture rich facial structures, which can be considered as a base deep model for facial representations. Since this pre-trained face model can represent face images for separating different identities, it cannot be used directly for depression analysis. Fig. 3 shows the detailed deep network architecture in our framework.

Next, in the fine-tuning step, the aim is to adapt the pre-trained network with depression data so that the network is capable of predicting the depression values given the input of image frames. Because depression estimation is considered as a regression problem, the loss function of the network for fine-tuning is changed to the Euclidean loss, which is appropriate for regression, other than using the softmax loss function in the pre-training step. Mathematically, the Euclidean loss function E computes the sum of squared differences of its two inputs, which can be written as:

$$E = \frac{1}{2N}\sum_{i=1}^{N} ||\hat{y}_i - y_i||^2, \tag{1}$$

where N is the number of samples, $\hat{y}_i$ is the output from the network and y_i is the ground truth.

Then the training image frames with their depression values are fed into the pre-trained network for fine-tuning. After this step, the network is capable of learning the depression representations given the input of image frames.

3.2 Dynamics-DCNN

In addition to appearance, we propose an architecture to model facial dynamics, by utilizing the optical flow between video frames. It is named dynamics-DCNN. Unlike the appearance model described above, the input of this model is formed by optical flows between several consecutive video frames. In this way, the input itself captures the motion between frames caused by the movement of the subject, which is then fed into the network to learn the motion representations.

Specifically, for each frame in the video, we compute the optical flow displacements between several consecutive frames. Since the depression recognition is focused on the face region in each video frame, the changes of face region between two frames are usually very subtle. Therefore, we compute optical flow between several

consecutive frames (e.g., every 10 frames), so that the motion of the face can be well captured and in the same time the video redundancy can be reduced.

The optical flow computed from each image is transformed into a "flow image" [34]. The three channels of the "flow image" are constructed by the horizontal and vertical components: x flow values, y flow values, as well as the flow magnitude. The values of each channel are then centered and normalized between 0 and 255, respectively. Fig. 4 shows some examples of RGB image frames and the generated flow images.

Given the "flow images" computed from video frames, a DCNN is trained for modeling the facial dynamics. The architecture and configurations remain largely the same as that used in the appearance DCNN, without the pre-training step. Fig. 2 shows an illustration of the two main networks.

3.3 Joint Tuning Layers

In our approach, the appearance-DCNN and the dynamics-DCNN are capable of predicting the depression values separately. In order to further improve the performance and integrate the two individual deep networks, we propose to construct joint tuning layers with a fine-tuning step, aiming at combining the appearance and dynamic models. Specifically, two fully connected layers are constructed with different numbers of hidden units (e.g., 512 and 256, respectively), connecting the concatenated feature layers (the FC layers) in both the appearance and dynamics networks. The final loss function still keeps the same Euclidean loss for regression task. The gradually decreasing number of hidden units in joint tuning layers, is designed to better converge for the single value regression. During training, the two DCNNs are trained separately, and then the final fine-tuning is conducted using the architecture with joint tuning layers, as shown in Fig. 2, where the input includes the RGB video frames and the computed "flow images".

4 Experiments

The experiments are conducted on two databases: the Audio/Visual Emotion Challenge (AVEC) 2013 [20] and 2014 [21] depression sub-challenge databases. In this section, we briefly describe the two databases first, and then show the experimental results. Finally we compare with other state-of-the-art methods.

4.1 AVEC2013 Depression Database

AVEC2013 depression database [20] is collected in the wild which contains 340 video clips from 292 subjects. A subset of the audio-visual depressive language corpus (AVid-Corpus) from AVEC2013 database is used for the depression sub-challenge. This subset contains video clips of subjects performing a human-computer interaction task, which is collected by a webcam as well as a microphone. There is only one subject in each video clips with no constrains when being recorded. The average length of each video clip is about 25 minutes. The age range of the subjects is from 18 to 63 years old, with an average age of 31.5 years. Some example images of this database are shown in Fig. 4 (top row). Specifically, for the depression sub-challenge, there are 150 videos from 82 subjects, split into three partitions: training, development, and test sets. Each of the three sets contains 50 video clips. For each video clip, a depression severity is assigned as the label, which was accessed using a standardized depression questionnaire, the Beck Depression Inventory-II (BDI-II) [20]. BDI-II scores range from 0 to 63, where 0-13 indicates none depression, 14-19 indicates mild depression, 20-28 indicates moderate, and 29-63 indicates sever depression. In our experiments, all data from training and development sets are used for training the deep models, while the test set is used to evaluate the overall performance for depression recognition.

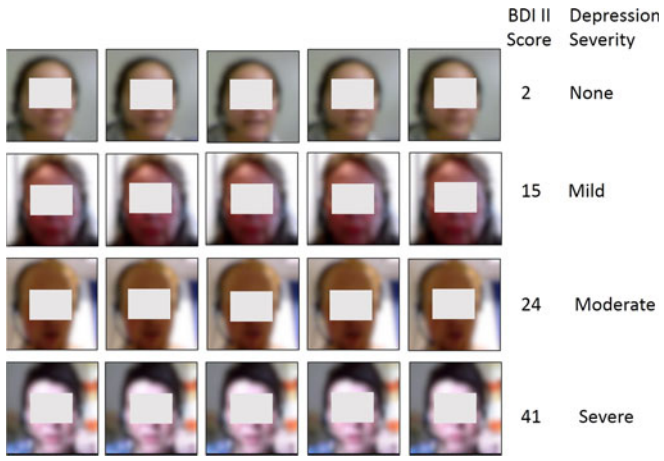

Fig. 1. Example video frames with depression value score (BSDII score) and depression severity categories from the AVEC2014 database. To protect the subjects, face images are blurred and eye regions are occluded. This kind of processing is applied to face images in other figures as well.

4.2 AVEC2014 Depression Database

AVEC2014 depression database was proposed for the Audio/Visual Emotion Challenge 2014 [21], where a subset of the audio-visual depressive language corpus (AViD-Corpus) is used for the depression sub-challenge. For the AVEC2014 challenge, two of the 12 tasks from AViD-Corpus are used, which are referred as Freeform and Northwind tasks. For both tasks, the recorded videos are split into three partitions: training, development, and test of 50 videos, respectively. In our experiments, we merge the training and development set from both Freeform and Northwind data as one training set. The overall performance is reported for video clips from the test set. Some example images of this database are shown in Fig. 1.

4.3 Experimental Settings

4.3.1 Face Region Detection and Alignment

In order to extract facial representations from the videos, the first step is to apply face detection and facial landmark localization for each video frame. In our experiments, we used the dlib [35]. Then within each video frame, the facial region is cropped and aligned by the eye locations with an image size of 256×256. This setting is kept for all face images for both training and testing.

4.3.2 Facial Dynamics Computation

After the above step, for each video clip in the dataset, a sequence of facial regions are extracted where faces are also aligned according to the eye locations. To compute the facial dynamics, we applied optical flow computation between two frames, with an interval of 10 frames, which is empirically selected and shows good performance in our experiments. A "flow image" is generated for each frame by taking the x and y flow values as the first and second channel. The third channel is created by calculating the magnitude of the optical flow. Those values are also centered around 128 and normalized between 0 and 255.

4.3.3 Subsampling

In order to reduce the large number of frames in each video clip, we applied a subsampling scheme that takes video frames with an interval of 100 frames for AVEC2013 and 10 frames for AVEC2014, respectively. In the original datasets, the videos in AVEC2013 are much longer than in AVEC2014. After the subsampling, there are about 380,000 video frames extracted from the AVEC2013 database, while about 50,000 for AVEC2014. The subsampling intervals are determined experimentally based on the number of frames in

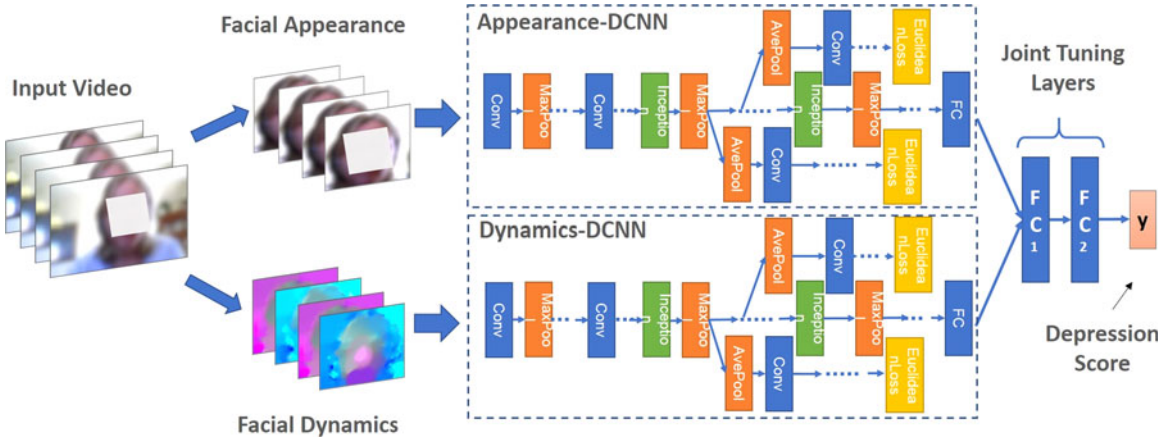

Fig. 2. Illustration of the proposed method for depression recognition using deep networks. The appearance DCNN takes facial images as input, while the dynamics DCNN uses the facial dynamics (e.g., optical flows) as input. The predicted depression score is computed by aggregating or averaging the individual predictions per frame from both DCNNs.

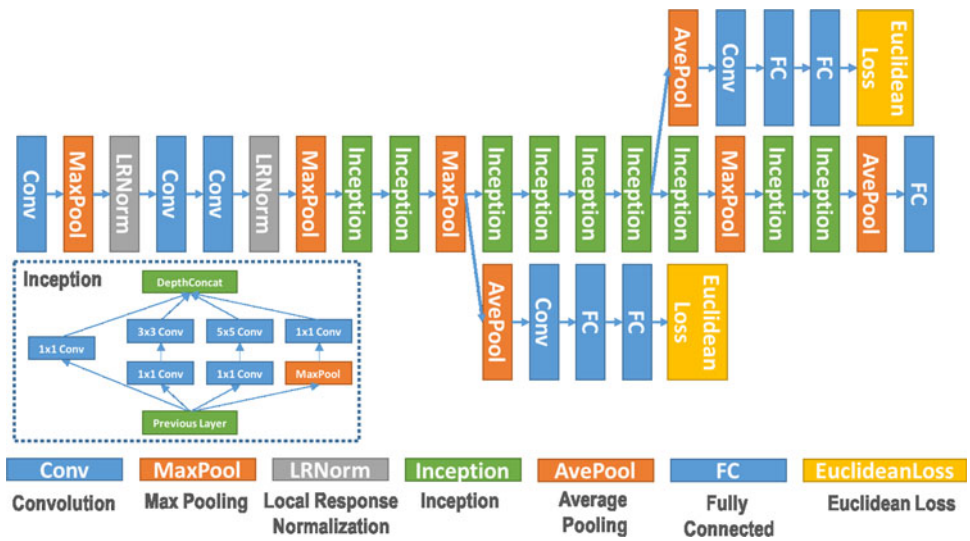

Fig. 3. The network architecture of the proposed method (composite networks) for depression recognition. The facial appearance and dynamics trainings are based on different pre-processing and pre-training, while using the same network structure for learning. Then the two networks are combined to form the final decision on depression diagnosis.

each video or the length of the videos, as well as the frequencies of facial variations in the videos. The subsamples are used as separate observations for training the deep networks. The predicted depression score is computed based on aggregating or averaging the predicted values from the (subsampled) frames in both appearance and dynamic models.

4.3.4 Deep Convolutional Neural Network

The layer configuration of our appearance and dynamics deep network is schematically shown in Fig. 2, and some details are described in Section 3. For appearance and dynamic DCNN, the model architectures are similar to the GoogLeNet model [32]. In our experiments, the joint tuning layers are designed as two fully connected layers with 512 and 256 hidden units, respectively. The networks are trained with stochastic gradient using caffe deep learning toolbox [32] with a batch size 32. In appearance DCNN, the model is fine-tuned from our pre-trained deep face model. This pre-trained model is trained on CASIA WebFace Database [33] with 494,414 images of 10,575 subjects. While in the dynamics DCNN, the training starts from scratch. The loss function is set to Euclidean loss for regression. The number of iterations for appearance model and dynamics model are set to 400,000 and 600,000, respectively. The selection of parameters is based on our experience in training deep networks, without any special tuning for the given dataset. The base learning rates are set to 0.001 and reduced by polynomial with gamma equals to 0.5. The momentum is set to 0.9 with weight decay equals to 0.0002. For joint tuning, the number of iterations is set to 200,000, with base learning rate 0.0001. All experiments are conducted using Titan-X GPU with 12 GB memory.

4.3.5 Performance Measure

For each video, the predicted depression value is computed by averaging the predicted values from both appearance model and dynamics model for each frame (subsampled) in the video. The

Fig. 4. Example image frames (top) and the generated flow images (bottom) from AVEC2014 dataset.

TABLE 2
Depression Recognition Results on AVEC2013 (Test Set)

Our Methods	RMSE	MAE
Facial Appearance Model	10.19	7.88
Facial Dynamics Model	10.02	7.87
Appearance & Dynamics (Ave.)	9.91	7.74
Appearance & Dynamics (Joint Tuning)	**9.82**	**7.58**

Ave. means score level fusion by taking average.

overall performance is measured using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).

The MAE is computed by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|. \tag{2}$$

And the RMSE is computed by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{3}$$

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

4.4 Performance of Individual Models

The results of depression recognition on AVEC2013 and AVEC2014 databases are shown in Tables 2 and 3, respectively. First, we explored the performance using individual deep model (appearance or dynamics) without any joint tuning procedure. From Table 2, one can see that, on AVEC2013, when only the appearance model is used, the MAE and RMSE achieved 7.88 and 10.19, respectively. While the MAE and RMSE are 7.87 and 10.02 when using the dynamics model, which is comparable to the appearance model. From Table 3 (AVEC 2014), when using appearance model, the MAE and RMSE are obtained 7.82 and 10.36, respectively. Comparable MAE and RMSE are also obtained (7.52 and 9.80 respectively) when using the dynamics model. This observation indicates that the dynamic information is important for depression diagnosis, and the dynamic DCNN can characterize the facial dynamics well. These results also show the effectiveness of appearance model as well as dynamics model, both of which are capable of learning the facial representations for depression recognition from video frames.

4.5 Overall Performance by Fusing the Individual Models

We also compute the performance by fusing the appearance and dynamics models. This fusion is conducted on the score level, which is computed by averaging the output values of depression severity from both the appearance and dynamic networks. The experimental results are shown in Tables 2 and 3 for AVEC2013 and AVEC2014, respectively. From Table 2, one can see that, the results after fusing the models obtained the MAE 7.74 and RMSE 9.91 on AVEC2013 database. On AVEC2014 (see Table 3), the fusion results achieved the MAE 7.53 and RMSE 9.73, respectively.

TABLE 3
Depression Recognition Results on AVEC2014 (Test Set)

Our Methods	RMSE	MAE
Facial Appearance Model	10.36	7.82
Facial Dynamics Model	9.80	7.52
Appearance & Dynamics (Ave.)	9.73	7.53
Appearance & Dynamics (Joint Tuning)	**9.55**	**7.47**

Ave. means score level fusion by taking average.

TABLE 4
Depression Recognition Result Comparison to Other Methods on AVEC2013 (Test Set)

Methods	RMSE	MAE
Baseline [20]	13.61	10.88
team-australia [8]	10.45	N/A
Brunel-Beihang [7]	11.19	9.14
Wen [19]	10.27	8.22
Our Method	**9.82**	**7.58**

Note that the listed results use video data only.

Both results perform better than those using each individual model. This observation shows that by fusing the appearance and dynamics models, the overall performance can be improved than using an individual model, which further implies the necessity of utilizing both facial appearance and dynamics for depression recognition.

4.6 Overall Performance by Joint Tuning

Next, we conduct experiments using the proposed joint tuning approach. The results are shown in the last row in Tables 2 and 3, for AVEC2013 and AVEC2014, respectively. It can be seen from the table that, when joint tuning is applied, the MAE and RMSE obtained are 7.58 and 9.82, respectively, on AVEC2013 database. These results are better than individual models. Further, the joint tuning results are also better than the score-level fusion (MAE 7.74 and RMSE 9.91) of the two models. Similar observations can also be found on AVEC2014 database (see Table 3), the best result is achieved by using the joint tuning, resulting in the MAE of 7.47, and RMSE of 9.55. These results show that the proposed joint tuning approach can better utilize both appearance and dynamics models, and the performance is improved. Moreover, the comparison with score-level fusion also shows the effectiveness of the proposed joint tuning approach.

4.7 Comparison with Pervious Methods

Finally, we compare our approach with other methods on both AVEC2013 and AVEC2014 databases. For a fair comparison, we show the results that are only using video data for depression recognition in Tables 4 and 5. From the table one can see that, our approach achieves better performance than the listed methods on both AVEC2013 and AVEC2014 databases. This further shows the effectiveness of our proposed approach for depression recognition.

5 Discussions and Conclusions

Since the AVEC2013 and AVEC2014 databases are used for the depression recognition competition, we also show our results with comparison to the competition results. Note that, our approach only utilized the video data without using audio cues, however in the listed competition results, audio based approaches are utilized in many of those methods. Combining audio cues could further

TABLE 5
Depression Recognition Result Comparison to Other Methods on AVEC2014 (Test Set)

Methods	RMSE	MAE
Baseline [21]	10.86	8.86
UUIMSidorov [36]	13.87	11.20
InaoeBuap [29]	11.91	9.35
Brunel [30]	10.50	8.44
BU-CMPE [31]	9.97	7.96
Our Method	**9.55**	**7.47**

Note that the listed results use video data only.

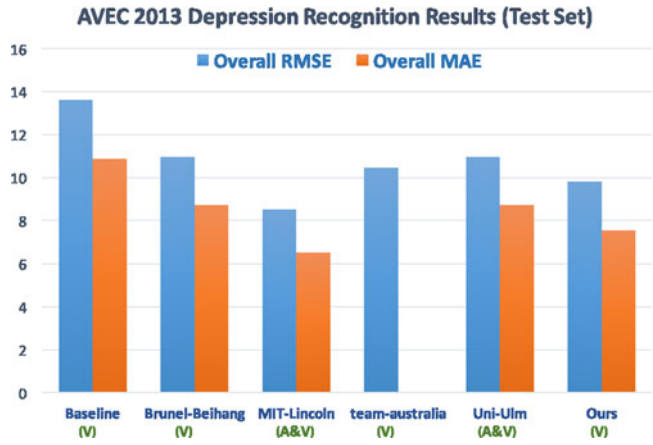

Fig. 5. Comparison of depression recognition results on AVEC2013 competition. Note that several of the listed methods utilize the audio data while our method only uses the visual data. (V) and (A) indicate the method utilizes video and audio data, respectively.

improve the performance, however, our focus here is exploring visual-based approaches for depression analysis.

The results of the AVEC2013 and AVEC2014 challenges are shown in Figs. 5 and 6. Note that, in these tables, most of the methods utilized both video and audio data for depression recognition, while in our approach, only video data are used. From Fig. 5, one can see that, our approach performs better than four methods on the AVEC2013 database, and comparable to the best results from [6], where both audio and video data are used. On the AVEC2014 database (see Fig. 6), our approach also achieved promising results, which is comparable to the top methods where both audio and video data are utilized. Top methods are those with lower bars in Figs. 5 and 6.

In summary, we have investigated the problem of depression diagnosis from video data. In order to model both facial appearance and dynamics for depression recognition, we proposed a new approach based on deep learning, which is for the first time to employ deep representations for depression analysis, to the best of our knowledge. In our proposed approach, a two-steam framework has been designed to take facial images and facial flows as input to model the depression information, called appearance and dynamics DCNN, respectively. Also, we have proposed to construct joint tuning layers, to combine the appearance and dynamics DCNN, which further improves the performance. Experimental results on two depression databases, AVEC2013 and AVEC2014, have shown that, our approach achieves better results compared to other visual based approaches for depression prediction.

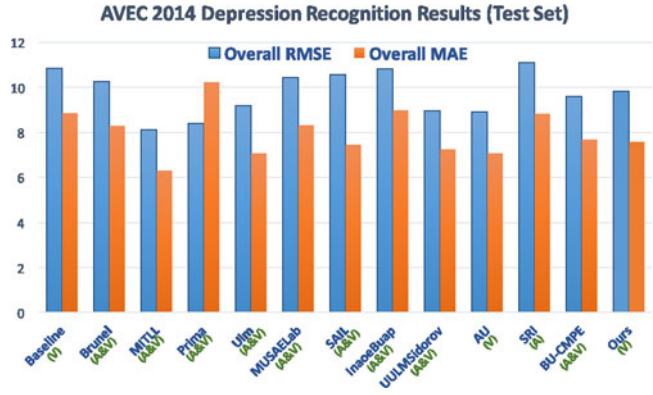

Fig. 6. Comparison of depression recognition results on AVEC2014 competition. Note that several of the listed methods utilize the audio data while our method only uses the visual data. (V) and (A) indicate the method utilizes video and audio data, respectively.

Moreover, our results using video data only can still achieve a comparable performance to the state-of-the-art approaches in the AVEC competition, where most methods utilized both video and audio data together.

Acknowledgments

This work is partially funded by Beijing Advanced Innovation Center for Imaging Technology. The authors thank the organizers of AVEC2013 and AVEC2014 for providing data for the study. The authors also thank the anonymous reviewers for valuable suggestions and comments to improve the paper. Guodong Guo is the corresponding author.

References

[1] R. Belmaker and G. Agam, "Major depressive disorder," *New England J. Med.*, vol. 358, no. 1, pp. 55–68, 2008.
[2] R. C. Kessler, et al., "The epidemiology of major depressive disorder: Results from the national comorbidity survey replication (NCS-R)," *JAMA*, vol. 289, no. 23, pp. 3095–3105, 2003.
[3] S. Salmans, *Depression: Questions You Have-Answers You Need*. Peoples Medical Society, Pennsylvania, USA, 1995.
[4] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 7, pp. 829–837, Jul. 2000.
[5] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, "Vocal acoustic biomarkers of depression severity and treatment response," *Biological Psychiatry*, vol. 72, no. 7, pp. 580–587, 2012.
[6] 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 *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 41–48.
[7] 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 *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 21–30.
[8] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: A multimodal approach," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 11–20.
[9] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting depression severity from vocal prosody," *IEEE Trans. Affective Comput.*, vol. 4, no. 2, pp. 142–150, Apr.-Jun. 2013.
[10] J. F. Cohn, et al., "Detecting depression from facial actions and vocal prosody," in *Proc. IEEE 3rd Int. Conf. Affective Comput. Intell. Interaction Workshops*, 2009, pp. 1–7.
[11] L. Chao, J. Tao, M. Yang, Y. Li, and J. Tao, "Multi task sequence learning for depression scale prediction from video," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, 2015, pp. 526–531.
[12] I. H. Jones and M. Pansa, "Some nonverbal aspects of depression and schizophrenia occurring during the interview." *J. Nervous Mental Disease*, vol. 167, no. 7, pp. 402–409, 1979.
[13] H. Ellgring, *Non-Verbal Communication in Depression*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
[14] R. L. Birdwhistell, "Toward analyzing american movement," *Nonverbal Commun.*, Oxford Univ. Press., pp. 134–143, 1974.
[15] 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 Vis. Comput.*, vol. 32, no. 10, pp. 641–647, 2014.
[16] N. Firth, "Computers diagnose depression from our body language," *New Scientist*, vol. 217, no. 2910, pp. 18–19, 2013.
[17] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, "Eye movement analysis for depression detection," in *20th IEEE Int. Conf. Image Process.*, 2013, pp. 4220–4224.
[18] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," *J. Personality Assessment*, vol. 67, no. 3, pp. 588–597, 1996.
[19] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inform. Forensics Secur.*, vol. 10, no. 7, pp. 1432–1441, 2015.
[20] M. Valstar, et al., "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Vis. Emotion Challenge*, 2013, pp. 3–10.
[21] M. Valstar, et al., "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 3–10.
[22] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.
[23] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.

[24] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image Video Retrieval*, 2007, pp. 401–408.
[25] H. Meng and N. Pears, "Descriptive temporal template features for visual motion recognition," *Pattern Recognit. Lett.*, vol. 30, no. 12, pp. 1049–1058, 2009.
[26] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
[27] S. De Jong, "Simpls: An alternative approach to partial least squares regression," *Chemometrics Intell. Laboratory Syst.*, vol. 18, no. 3, pp. 251–263, 1993.
[28] T. R. Almaev and M. F. Valstar, "Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 356–361.
[29] 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," in *Proc. ACM 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 49–55.
[30] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, and S. Turabzadeh, "Automatic depression scale prediction using facial expression dynamics and regression," in *Proc. ACM 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 73–80.
[31] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. ACM 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 19–26.
[32] C. Szegedy, et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[33] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, http://arxiv.org/abs/1411.7923
[34] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. Comput. Vision-ECCV*, 2004, pp. 25–36.
[35] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learning Res.*, vol. 10, pp. 1755–1758, 2009.
[36] M. Sidorov and W. Minker, "Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach," in *Proc. 4th Int. Workshop Audio/Vis. Emotion Challenge*, 2014, pp. 81–86.