Automatic Depression Detection via Learning and Fusing Features From Visual Cues

Yanrong Guo®, Chenyang Zhu, Shijie Hao®, and Richang Hong®, Senior Member, IEEE

Abstract—Depression is one of the most prevalent mental disorders, which seriously affects one’s life. Traditional depression diagnostics commonly depend on rating with scales, which can be labor-intensive and subjective. In this context, automatic depression detection (ADD), aiming to assist medical experts in their diagnosis and analysis, has been attracting more attention for its better objectivity and fewer laborious interventions. A typical ADD model detects depression via automatically extracting task-specific features from medical records, such as video sequences, and sending them into a classifier for assistive prediction. However, it remains challenging to effectively extract depression-specific information from long sequences, thereby hindering a satisfying accuracy. In this article, we propose a novel ADD method via learning and fusing features from visual cues. Specifically, we first construct temporal dilated convolutional network (TDCN), in which multiple dilated convolution blocks (DCBs) are designed and stacked, to learn the long-range temporal information from sequences. Then, the featurewise attention (FWA) module is adopted to fuse different features extracted from TDCNs. The module learns to assign weights for the feature channels, aiming to better incorporate different kinds of visual features and further enhance the detection accuracy. Our method achieves the state-of-the-art performance on the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC_WOZ) dataset compared with other visual-feature-based methods, showing its effectiveness.

Index Terms—Depression detection, dilated convolution, fusion, visual cue.

I. INTRODUCTION

DEPRESSION is a major mental disorder in nowadays society. It has a negative impact on one’s life, even leading to suicide in extreme conditions [1]. What is more urgent, despite that depression has become one of the most prevalent psychiatric disorders, the number of people suffering from this disease is still on the rise. Luckily, the symptoms of depression can be alleviated on the premise of in-time treatment. Therefore, accurate depression diagnosis has become the key factor. Traditional depression diagnostics mostly rely on subjective rating with scales, such as the Eight-item Patient Health Questionnaire depression scale (PHQ-8) [2], where patients with scores no smaller than 10 are considered as depressed. It needs the participation of experienced psychologists, which can be labor-intensive. However, subjective bias may still exist, therefore leading to misdiagnosis. Besides, different from some physical diseases, the severity of mental disorders may vary at different time [3], [4], which brings extra challenges to the diagnostics. Consequently, the technique named automatic depression detection (ADD) has been drawing increasing attention. The goal of building ADD models is to assist medical expert in judging the mental status from large amount of recorded data, such as interviewing videos. A typical ADD model is built by following the pattern classification paradigm, in which discriminant features are automatically learned from the massive input data, and then sent into a predictor to produce labels or scores, which are used as the reference information for medical experts. From the above process, we can see that an ADD system has the potential of helping medical experts, as it is more objective and reproducible and requires much fewer laborious interventions.

Images or videos provide rich visual cues of mental conditions. Here, the visual cues in our application mainly refer to the facial appearance variations or the head movements of an interviewee. Compared with common people, patients suffering from depression usually have different facial actions, such as glazed expressions and abnormal head movements [5], [6]. Therefore, visual data have become a main information source for detecting depression. However, depression-related visual cues do not appear as short-time behaviors. Instead, it is not distinguishable until a relatively long observation time is given. Thus, the detection model usually requires a large perceptive field to explore depression-specific information. Therefore, as for depression detection based on visual cues, it is necessary to take care of the dynamic characteristics of data, which can be called temporal information. Some works try to solve this issue with long short-term memory (LSTM) [7] or temporal convolutional network (TCN) [8] and have achieved promising results. Nevertheless, limitations still exist in two aspects. On one hand, these networks are less suitable for processing the overlong sequences. On the other hand, they prefer to using a single visual cue, ignoring the complementarity between different kinds of visual information.

To tackle these limitations, we propose a novel ADD method based on visual cues in this article. The whole
framework of our method is shown in Fig. 1, which has two main modules: temporal dilated convolution network (TDCN) and featurewise attention (FWA). Specifically, TDCN extracts depression-specific information by fully using the dilated convolutions of various receptive fields. FWA further enhances the feature representation ability by assigning different weights to feature channels. We achieve the state-of-the-art performance compared with other visual-based works, e.g., 0.800 F1-score, on the validation set of the Distress Analysis Interview Corpus Wizard-of-Oz (DAIC_WOZ) dataset. Generally, the contributions of our work are twofold.

1) We propose TDCN to effectively extract the temporal information from long video sequences. Within TDCN, two parallel dilated convolutional modules are applied to learn useful temporal information for detecting depression, and the max pooling layers are adopted to solve the overlength problem.

2) We construct the FWA module to fuse the learned features from different TDCN branches. With the attention module, our method is able to further highlight the important part of the learned features, thereby enhancing the ADD accuracy.

II. RELATED WORK

Many methods have been developed for the task of ADD. Some methods are based on single modality such as audio [9], [10], [11], text [12], [13] or visual cues [7], [14], while others combine at least two modalities [15], [16]. Zhang et al. [17] propose a self-supervised method to extract the audio embedding and achieve great improvement on the depression detection task. Niu et al. [16] propose a hierarchical graph attention network for depression detection, which considers the interrelation between text and audio modalities. Apart from these modalities acquired from interviews, there are also some works using other sensors to detect depression. Lu et al. [18] propose a new method for ADD that uses a skeletal representation of the human body based on gait data. Shen et al. [19] use multichannel electroencephalogram (EEG) signals to detect depression.

Early ADD works using visual cues typical follow the traditional roadmap of a pattern recognition task, i.e., handcrafted features are first extracted based on visual data, and then novel classifiers are constructed to achieve depression detection. Either of these two steps can be the key factor in ADD research. For example, Wang et al. [20] design probabilistic classifiers for the purpose of judging the mental state of candidates through the variations in their facial expressions. Yang et al. [21] propose a gender-specific decision tree for depression detection. As for the visual cues, geometric features from 2-D facial landmarks are extracted. Besides, gaze and pose features and emotion evidence measures provided by AVEC2016 [22] are also added. Despite the obtained promising results, they take the average of each hand-crafted feature as input without considering the long-term correlations, inevitably losing the temporal information. Some other methods also have a similar issue. Pampouchidou et al. [23] extract low-level features such as Landmark Motion History Images (LMHI) and Landmark Motion Magnitude (LMM) from 2-D facial landmarks. However, the temporal information of other visual cues is mostly unused. In a word, works based on the traditional pattern recognition paradigm have a common problem, that is, the feature extraction processes are usually constructed based on some prior knowledge or ad hoc rules, which can be less comprehensive and robust for the ADD task.

Recently, more and more ADD methods based on the deep learning models have been proposed, as they are advantageous for solving the issue of hand-crafted features. To learn the correlation and dynamic variation between frames, the LSTM neural networks [24] or other recurrent neural networks (RNNs) for such temporal data can be usually adopted. For example, both [25] and [26] choose LSTM as their backbone network. However, the performance of LSTM degrades when
sequences become too long [27]. In other words, LSTM is not good at handling overlength sequences in practice. Besides, the number of parameters becomes huge when LSTM receives overlong sequences, possibly leading to speed reduction and overfitting in the training process. One feasible solution is to shorten the input length directly by sampling. Wang et al. [7] introduce sampling slicing LSTM multiple instance learning (SS-LSTM-MIL), a method based on multiple instance learning with 2-D facial landmarks as input. Its sampling strategy is that only patient-speaking frames are chosen. It outperforms other visual-based methods in the ADD task. Another solution is to do embedding first to unify the length of sequences or feature dimensions. Du et al. [14] propose DepArt-Net for depression recognition, which encodes four visual cues into long-term representations and then applies atrous residual TCNs with attentive temporal pooling. Haque et al. [8] leverage a casual convolutional neural network (C-CNN), which is a TCN [28] actually, with a sentence-level summary embedding to detect the major depressive disorder. However, TCN still has some shortages that are analyzed in Section III.

Different from the above methods, our work is designed to better relieve the overlength issue. First, to decrease the computational complexity of our model, we split samples into several sequences with fixed size. Second, we propose the TDCN model which fully uses the information from the perceptive fields of various scales.

III. METHOD

A. Overview

Generally, as shown in Fig. 1, the overall framework of our method includes four parts: data preprocessing, feature extraction, fusion, and classification. Specifically, we first divide samples into sequences with fixed shape. Then, we use two TDCN branches to learn discriminate task-specific features. Finally, as the main part of the whole model, the details of TDCN and FWA are described in the following.

B. Temporal Dilated Convolution Network

Fig. 1 shows the general structure of the proposed TDCN. Generally, TDCN is a multilayer structure, consisting of five dilated convolutional blocks (DCBs) and four max pooling layers. On one hand, within a certain TDCN layer, DCB explores the useful information at different perceptive ranges. On the other hand, along the TDCN pipeline, the max pooling layers keep on shrinking the feature resolutions and gradually extract the most important responses. Therefore, TDCN can be seen as a feature learning module that extracts depression-specific information from multiple scales.

DCB is the main component of a TDCN module, which is described in the following. As shown in Fig. 2, to enlarge the receptive field, we construct two parallel dilated convolution paths. Each path contains three convolutional networks with different dilation factors for a multiscale receptive field. Given an input $X = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T \times D}$ where $T$ is the time steps and $D$ is the feature dimension, the operation of dilated convolution can be expressed as

$$\mathcal{F}(i) = \sum_{i=0}^{k-1} \text{filter}(i) \cdot x_{i+d(i-1)} + b$$

(1)

where $d$ is the dilation factor, $k$ is the kernel size, and $b$ is the bias. Zero padding is adopted to keep the same shape of its input and output. As for our ADD application, the dilation factor doubles by a factor of 2, aiming to obtain temporal information from different spans along the timeline. At different dilation factors, the two paths join together by means of the summation and activation operations. We choose exponential linear unit (ELU) [29] as DCB’s activation function, which is defined as

$$f_{elu}(x) = \begin{cases} x & \text{if } x \geq 0 \\ e^x - 1 & \text{if } x < 0. \end{cases}$$

(2)

We add a residual block [30] in our DCB to avoid the degradation issue when the network goes deeper. In addition, to keep the same shape of elementwise addition in shortcut connections, we add a 1-D convolutional layer with a kernel size of 1 in every DCB. At the end of DCB, the batch normalization layer is applied to accelerate the training process. Besides, it is able to alleviate the problem of gradient vanishing. Of note, we remove it from the last DCB in TDCN to reserve the distribution of different features.

After each DCB, we add a max pooling layer except the last one in TDCN. With this operation, its output tensor obtains a broader receptive field. In another word, the pooling layers gradually aggregate the most important information from long sequence. In addition, the max pooling layers also cut down the length of sequences and retain the important part, reducing the complexity of our model.

In the following, we further analyze the difference between TCNs [27] and our TDCN module. To enlarge the receptive field of the traditional convolution, Bai et al. [28] develop a
new architecture for TCN using dilated convolutions. However, as a sequence-to-sequence architecture, TCN is less suitable for handling overlong video sequences due to insufficient receptive field. Take Fig. 3(a) for example, given an input \( X = [x_1; x_2, \ldots, x_T] \in \mathbb{R}^{T \times D_{in}} \), the length of TCN’s output \( Y = [y_1; y_2, \ldots, y_T] \in \mathbb{R}^{T \times D_{out}} \) is the same as \( X \). Therefore, the size of \( Y \) can be large in applications when the input is also very long. In this context, \( y_T \) in \( Y \) can be selected as the input of the final classifier. The rationale comes from the use of causal convolutions, such as the connection between units of different layers in Fig. 3(a). Based on this construction, the network is expected to be deeper when a larger receptive field is needed in our ADD application. This potentially brings in the side effects such as oversize dilation factor and more computational cost, which may lower the detection accuracy of the TCN-based model, as well as the efficiency.

Our TDCN relieves the above limitation using a different structure. Fig. 3(b) shows the \( n \)th DCB+max pooling layer in our TDCN. In the DCB part, we incorporate multiple dilated convolutions with increasing dilated factors (\( d = 1, 2, 4 \)). Then a max pooling layer is followed, shrinking the feature length and naturally enlarging the receptive field for its neighboring \((n + 1)\)th DCB, of which the input size has been reduced into \( T/2M \) (\( M = 2^{n-1}, n = 1, \ldots, 5 \)). Compared with the TCN structure, TDCN better controls the receptive field of long sequences in terms of flexibility and efficiency. This is also empirically validated by the experimental results.

\[ s_j = \frac{1}{T} \sum_{i=0}^{T-1} X_{i,j} \]  
\[ h = f_{att}(s, W) = \sigma_{sigmoid}(W_2(f_{relu}(W_1s))) \]  
\[ \tilde{X} = f_{scale}(x, h) = X \odot \tilde{H} \]  
\[ X \in \mathbb{R}^{T \times 2D} \]  
\[ s \in \mathbb{R}^{2D} \]  
\[ \text{concatenate} \]  
\[ \text{Global Average Pooling} \]  
\[ \text{Linear Layer} \]  
\[ \text{ReLU} \]  
\[ \text{Sigmoid} \]  

**Fig. 4.** Structure of the FWA module.

**C. Featurewise Attention**

In the visual-based ADD task, various visual cues represent the interviewee’s status from different views. For example, the sequence of landmark positions represents one’s facial morphing, while the pose vector sequence represents how one’s face orientates during the interviewing process. They all potentially contribute to the final inferring process in our ADD research. To make the most of visual information, we construct the FWA module, inspired from [31], to effectively combine the different features. We first concatenate the features learned from different TDCN branches directly, obtaining \( X \in \mathbb{R}^{T \times kD} \), where \( D \) is the number of feature dimension and \( k \) is the number of TDCN branches. They are then fed into the FWA module to learn the weight of each feature dimension based on its significance. The details of the FWA module can be seen in Fig. 4. The global average pooling is applied to achieve a featurewise vector \( s \in \mathbb{R}^{kD} \), obtained as

**IV. EXPERIMENTS**

**A. Dataset and Implementation Details**

In our research, we use the DAIC_WOZ dataset [32] to evaluate the effectiveness of our model. The DAIC collects interviews of semistructured clinical. The Wizard-of-Oz interviews
use a virtual avatar named Ellie to interact with patients, asking fundamental questions, and collecting audio, video, and depth sensor recordings meanwhile. In general, this dataset includes audio and video recordings and text transcripts from audio data. The number of samples for training/validation/testing is 107/35/47, respectively. As our model needs inputs with the same shape, it is necessary to resample the sequences and set them into a unified length. Meanwhile, the overall length of the input is expected to be relatively long so as to preserve the temporal information. To this end, we empirically perform the head-first resample to all the sequences in the dataset and set the length of every sample as 5000.

To protect the privacy of participants, DAIC_WOZ does not release the raw video recordings. Instead, several visual cues extracted from video recordings with OpenFace toolkit [37] are provided, including 68 2-D/3-D facial landmarks, action units (AUs), gaze vectors, head-pose vectors, and histogram of oriented gradients (HOG) features. These cues can be seen as low-level features and are assumed to contain the task-specific information. 2-D facial landmarks record the 2-D coordinates of 68 facial feature points with a total of 136 dimensions. The pose features contain six dimensions, including X, Y, Z, Rx, Ry, and Rz, where X, Y, and Z are the position coordinates of the head, and Rx, Ry, and Rz are the head rotation coordinates. The gaze vectors record the eye-gaze directions. AUs record the existence of 20 AUs. Based on the proposed TDCN, discriminative features that are used in the depression detection task can be learned from these cues. Of note, HOG is a feature descriptor used for object detection, of which the size is too large for our application. As for the remaining visual cues, we select two of them to validate the effectiveness of the FWA module, i.e., 2-D facial landmarks and head-pose features. Of note, we empirically find that incorporating three or more kinds of visual cues does not gain extra detection accuracy, which would be discussed in the analysis of our method.

The implementation details are described in the following. Depressed subjects are labeled as 1 (being positive), and nondepressed subjects are labeled as 0 (being negative). The feature dimensions of DCBs are 256, 256, 128, 64, 64 for 2-D facial landmark features and 128, 64, 256, 128, 64 for the head-pose features. During the training stage, we choose SGD as the optimizer whose learning rate is 2e−3 and the momentum rate is 0.9. The minibatch size is set as 8.

B. Evaluation Metrics

As shown in (6), we choose the metrics commonly used in classification tasks, i.e., accuracy, precision, recall, and F1-score, to evaluate performance. TP, TN, FP, and FN denote true positive, true negative, false positive, and false negative, respectively. As for our application, for example, TP means the number of depressed subjects that are predicted as depressed, and TN means the number of nondepressed subjects that are predicted as nondepressed

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

\[
\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

Based on these, accuracy represents the percentage of correctly classified samples in the total number of samples. Recall represents the percentage of positive samples that are correctly predicted. Precision represents the percentage of correctly predicted positive subjects in the subjects predicted as positive. F1-score jointly considers the metrics of precision and recall. In particular, recall is clinically important, as it reflects a model’s capability of finding out patients suffering from depression from all the participants. In another word, the price that the detection model misses one depressed subject is much higher. In addition, F1-score also matters in our application as this metric jointly considers recall and precision.

C. Comparison With Other Methods

We compare our method with the related ADD methods using the visual modality. Considering that most works evaluate their methods on the validation set, we split a part of the original training set for tuning the hyperparameters. Then we report our results on the validation set for a fair comparison. The results are shown in Table I, in which “A,” “V,” and “L” denote audio, visual, and textual modalities, respectively.

From Table I, we can see several general trends. On one hand, the overall results demonstrate the usefulness of the visual cues in the ADD task, both for the multimodality and single-modality methods. On the other hand, two factors contribute to the effectiveness of an ADD model, that is, the subtle construction of learning models and the effective incorporation of features. As for the comparison between the related methods and ours, we have the following detailed observations. First, we achieve very competitive performance among all the methods. Our method achieves the best performance in the metrics of accuracy, recall, and F1-score and achieves the second highest score on the precision value. Second, the results of our method surpass the ones based on multimodalities, demonstrating that our method is able to fully explore the discriminative information for accurate depression detection. The better performance of our model in the quantitative comparison attributes to the following two reasons. Compared with C-CNN [8] and SS-LSTM-MIL [7], the proposed TDCN is more suitable for the depression detection task than TCN and LSTM, as TDCN keeps more attentions on enlarging the receptive fields in various aspects. In addition, we introduce an effective fusion module into our framework, while methods such as Topic [36] and C-CNN [8] only adopt simple feature concatenation.

D. Analysis of Our Method

We conduct more experiments to investigate the influence and the effectiveness of our method’s elements. Of note, in these experiments, we follow the original partition of the dataset (107/35/47) and tune hyperparameters with which our model performs best on the original validation set. Consequently, the results on the validation set are different from those in Table I.
We first show the necessity of incorporating different kinds of visual cues. As shown in Table II, we compare the result of our model with its intermediate version, which only uses one single cue with one TDCN branch. We can observe a general trend that the performance of incorporating two kinds of visual cues is better than the performance based on a single kind of visual cue. Specifically, the recall metric is largely improved in both the validation set and the test set. In another word, our model is able to detect more depressed individuals from all the subjects, which can be meaningful in clinical applications. In contrast, compared with the performance of accuracy and precision, the recall scores of the single-cue models are relatively low, especially for the results of the test set, which makes them almost useless in clinical applications. These results again empirically validate the effectiveness of incorporating different kinds of visual cues in the ADD task.

As a step further, as shown in Table III, we report the results of different combinations of visual cues. In general, compared with the results in Table II, we can see that almost all the combinations obtain better performance than the single-cue configurations. Specifically, we observe that incorporating the landmark feature is important to the final results, especially for the metrics of recall and F1-score. The reason is that the facial landmarks provide more details of how one’s facial organs morph in a fine scale. Based on that, we empirically find that the 2-D Landmarks+Pose configuration achieves the best performance, which we adopt in our research. As for our application, we note that it might not be the best choice to use all kinds of visual cues. From Table III, for example, we can see that the performance of incorporating three and four kinds of visual cues has an obvious decline. The reason is that the model falls into the overfitting problem when facing more features [38]. Specifically, a new TDCN branch is needed when adding another kind of visual cue, bringing in a significant increase in the parameter size for the whole model. As a result, the overfitting problem is inevitable given that the number of training subjects is still limited.

We also investigate the effects of different resampling strategies, as shown in Table IV. In the data preprocessing stage, we empirically resample the sequence into a fixed length for all the training and test subjects. To represent a subject, head-first means that we only use a part of the sequence from the beginning. As for the average strategy, the sequence of a subject is divided into pieces with the fixed length. The results based on this strategy are obtained through taking the average of the soft predicting scores of these pieces. From Table IV, we can see that the head-first sampling strategy consistently performs better than the average strategy in both the sets. As the interviewing process is relatively long, many divided subsequences of one subject may not be depression-specific. In this context, the results based on the average strategy can be inevitably lowered in a statistical sense.

In the following, we conduct several ablation studies on the proposed model in terms of the backbone TDCN, the FWA module, and the max pooling operation. First, we replace TDCNs with TCNs in our model. The results in Table V show...
that TDCN significantly outperforms TCN in all the evaluation metrics. The results empirically validate the effectiveness of TDCN in aggregating global temporal information. Besides, our TDCN has much lower FLOPs (4.5G versus 10.7G, the inputs are 2-D Landmark sequences), showing its computational efficiency.

To show the significance of the FW A module, we construct a degraded model without FWA. In another word, the features learned from TDCN branches are simply concatenated and as well as the effectiveness of the elements in the proposed method.

As for the future work, we plan to include other modalities, e.g., text and audio, into our framework to enhance the ADD system’s performance. The main challenge here is to learn a common feature representation that facilitates the effective fusion of the heterogeneous input data. Furthermore, we plan to enhance the computational efficiency of the model with quantization techniques on the premise of preserving detection accuracy, aiming to make our model more practical in real-world applications.

### TABLE V

| Backbone | Validation Set | Test Set |
|----------|----------------|---------|
| TDCN     | 0.857 0.917 0.733 0.815 | 0.586 0.500 0.546 0.522 |
| TCN      | 0.686 0.583 0.709 0.636 | 0.596 0.286 0.308 0.296 |

### TABLE VI

| Model               | Accuracy | Recall | Precision | F1-score |
|---------------------|----------|--------|-----------|----------|
| With FWA            | 0.857    | 0.917  | 0.733     | 0.815    |
| Without FWA         | 0.771    | 0.583  | 0.709     | 0.636    |

### TABLE VII

| Pooling Operation  | Validation Set | Test Set |
|--------------------|----------------|---------|
| Max Pooling        | 0.857 0.917 0.733 0.815 | 0.575 0.143 0.200 0.167 |
| Average Pooling    | 0.857 0.917 0.733 0.815 | 0.600 0.643 0.450 0.530 |

### REFERENCES

[1] F. Hao, G. Pang, Y. Wu, Z. Pi, L. Xia, and G. Min, “Providing appropriate social support to prevention of depression for highly anxious sufferers,” IEEE Trans. Computat. Social Syst., vol. 6, no. 5, pp. 879–887, Oct. 2019.

[2] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, “The PHQ-8 as a measure of current depression in the general population,” J. Affect. Disorders, vol. 114, nos. 1–3, pp. 163–173, Apr. 2009.

[3] A. E. Kaczynski and S. L. Blase, “Rebooting psychotherapy research and practice to reduce the burden of mental illness,” Perspect. Psychol. Sci., vol. 6, no. 1, pp. 21–37, Jan. 2011.

[4] A. J. Mitchell, A. Vaze, and S. Rao, “Clinical diagnosis of depression in primary care: A meta-analysis,” Lancet, vol. 374, no. 9690, pp. 609–619, Aug. 2009.

[5] P. Caligiuri and J. Ellwanger, “Motor and cognitive aspects of motor retardation in depression,” J. Affect. Disorders, vol. 57, nos. 1–3, pp. 83–93, Jan. 2000.

[6] G. Parker, D. Hadzi-Pavlovic, H. Brodaty, P. Boyce, and K. Eyer, “Psychomotor disturbance in depression: Defining the constructs,” J. Affect. Disorders, vol. 27, no. 4, p. 255, 1993.

[7] Y. Wang et al., “Automatic depression detection via facial expressions using multiple instance learning,” in Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI), Apr. 2020, pp. 1933–1936.

[8] A. Haque, M. Guo, A. S. Miner, and L. Fei-Fei, “Measuring depression symptom severity from spoken language and 3D facial expressions,” 2018, arXiv:1811.08592.

[9] M. Muzammel, H. Salam, Y. Hoffmann, M. Chetouani, and A. Othmani, “AudiVowelConsNet: A phoneme-level based deep CNN architecture for clinical depression diagnosis,” Mach. Learn. Appl., vol. 2, Dec. 2020, Art. no. 100005.

[10] S. Saidi, S. B. Othman, and S. B. Saoud, “Hybrid CNN-SVM classifier for efficient depression detection system,” in Proc. 4th Int. Conf. Adv. Syst. Emergent Technol. (IC_ASET), Dec. 2020, pp. 229–234.

[11] L. He et al., “Deep learning for depression recognition with audiovisual cues: A review,” Inf. Fusion, vol. 80, pp. 56–86, Apr. 2022.

[12] D. Xezonaki, G. Paraskopoulos, A. Potamianos, and S. Sarayanan, “Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews,” 2020, arXiv:2006.08336.

[13] H. Dinkel, M. Wu, and K. Yu, “Text-based depression detection on sparse data,” 2019, arXiv:1904.05154.

[14] Z. Du, W. Li, D. Huang, and Y. Wang, “Encoding visual behaviors with attentive temporal convolution for depression prediction,” in Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2019, pp. 1–7.

[15] H. Soliman and E. A. Pustozarov, “The detection of depression using multimodal models based on text and voice quality features,” in Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng. (ElCon-Rus), Jan. 2021, pp. 1843–1848.
[16] M. Niu, K. Chen, Q. Chen, and L. Yang, “HCAG: A hierarchical context-aware graph attention model for depression detection,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 4235–4239.

[17] P. Zhang, M. Wu, H. Dinkel, and K. Yu, “DEPA: Self-supervised audio embedding for depression detection,” in Proc. 29th ACM Int. Conf. Multimedia, New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 135–143.

[18] H. Lu, W. Shao, E. Ngai, X. Hu, and B. Hu, “A new skeletal representation based on gait for depression detection,” in Proc. IEEE Int. Conf. E-Health Netw., Appl. Services (HEALTHCOM), Mar. 2021, pp. 1–6.

[19] J. Shen et al., “An optimal channel selection for EEG-based depression detection via kernel-target alignment,” IEEE J. Biomed. Health Inform., vol. 25, no. 7, pp. 2545–2556, Jul. 2021.

[20] P. Wang et al., “Automated video-based facial expression analysis of neuropsychiatric disorders,” J. Neurosci. Methods, vol. 168, no. 1, pp. 224–238, Feb. 2008.

[21] L. Yang, D. Jiang, L. He, E. Pei, M. C. Ovaneke, and H. Sahli, “Decision tree based depression classification from audio and language information,” in Proc. 6th Int. Workshop Audio/Visual Emotion Challenge, Oct. 2016, pp. 89–96.

[22] M. Valstar et al., “AVEC 2016: Depression, mood, and emotion recognition workshop and challenge,” in Proc. 6th Int. Workshop Audio/Visual Emotion Challenge, Oct. 2016, pp. 3–10.

[23] A. Pampouchidou et al., “Depression assessment by fusing high and low level features from audio, video, and text,” in Proc. 6th Int. Workshop Audio/Visual Emotion Challenge, Oct. 2016, pp. 27–34.

[24] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[25] T. Al Hanai, M. Ghassemi, and J. Glass, “Detecting depression with audio/text sequence modeling of interviews,” in Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), Sep. 2018, pp. 1716–1720.

[26] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, “Multi-level attention network using text, audio and video for depression prediction,” in Proc. 9th Int. Audio/Visual Emotion Challenge Workshop, 2019, pp. 81–88.

[27] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, “Temporal convolutional networks: A unified approach to action segmentation,” in Proc. 14th Eur. Conf. Comput. Vis. (ECCV), 2016, pp. 47–54.

[28] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” 2018, arXiv:1803.01271.

[29] D. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” in Proc. 4th Int. Conf. Learn. Represent. (ICLR), San Juan, PR, USA, May 2016, pp. 1–14.

[30] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 630–645.

[31] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[32] J. Gratch, R. Arstein, G. Lucas, G. Stratou, and L. P. Morency, “The distress analysis interview corpus of human and computer interviews,” in Proc. Lang. Resour. Eval. Conf., 2014, pp. 3123–3128.

[33] J. R. Williamson et al., “Detecting depression using vocal, facial and semantic communication cues,” in Proc. 6th Int. Workshop Audio/Visual Emotion Challenge, 2016, pp. 11–18.

[34] S. Song, L. Shen, and M. Valstar, “Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features,” in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2018, pp. 158–165.

[35] M. Nasir, A. Jati, P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou, “Multimodal and multiresolution depression detection from speech and facial landmark features,” in Proc. 6th Int. Workshop Audio/Visual Emotion Challenge, Oct. 2016, pp. 43–50.

[36] Y. Gong and C. Poellabauer, “Topic modeling based multi-modal depression detection,” in Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge, Oct. 2017, pp. 69–76.

[37] T. Baltrusaitis, P. Robinson, and L.-P. Morency, “OpenFace: An open source facial behavior analysis toolkit,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2016, pp. 1–10.

[38] W. Wang, D. Tran, and M. Feiszli, “What makes training multi-modal classification networks hard?” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12695–12705.

Yanrong Guo received the Ph.D. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2013. She was a Post-Doctoral Researcher with The University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC, USA, from 2013 to 2016. She is currently a Professor with the School of Computer and Information, HFUT, where she is also with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education. Her research interests include pattern recognition and medical image analysis.

Chenyang Zhu received the B.E. degree from the Hefei University of Technology, Hefei, China, in 2020, where he is currently pursuing the master’s degree with the School of Computer Science and Information Engineering. His current research interests include pattern recognition.

Shijie Hao received the Ph.D. degree from the Hefei University of Technology (HFUT), Hefei, China, in 2012. He is currently an Associate Professor with the School of Computer Science and Information Engineering, HFUT, where he is also with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education. His research interests include image processing and pattern recognition.

Richang Hong (Senior Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Fellow with the School of Computing, National University of Singapore, Singapore, from 2008 to 2010. He is currently a Professor with the Hefei University of Technology, Hefei, where he is also with the Key Laboratory of Knowledge Engineering with Big Data, Ministry of Education. He has coauthored over 70 publications in the areas of his research interests, which include multimedia content analysis and social media.

Dr. Hong is a member of the ACM and the Executive Committee member of the ACM SIGMM China Chapter. He was a recipient of the Best Paper Award from the ACM Multimedia 2010, the Best Paper Award from the ACM ICMR 2015, and the Honorable Mention of the IEEE TRANSACTIONS ON MULTIMEDIA Best Paper Award. He has served as the Technical Program Chair for the MMM 2016. He has served as an Associate Editor for the IEEE Multimedia Magazine, Neural Processing Letter (Springer), Information Sciences (Elsevier), and Signal Processing (Elsevier).