ORIGINAL ARTICLE

Journal Section

Intelligent System for Depression Scale Estimation with Facial Expressions and Case Study in Industrial Intelligence

Lang He[1*] | **Chenguang Guo** [2*†] | **Prayag Tiwari**[3*] | **Hari Mohan Pandey** [4†] | **Wei Dang**[5]

[1]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi, China

[1]Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi, China

[2]School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China

[3]Department of Computer Science, Aalto University, Finland

[4]Department of Computer Science, Edge Hill University, Ormskirk, United Kingdom

[5]Shaanxi Mental Health Center, Xi'an Shaanxi, China

Correspondence
Chenguang Guo, School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710072, China.
Hari Mohan Pandey, Department of Computer Science, Edge Hill University, Ormskirk, UK
Email: guochg@nwpu.edu.cn; Pandeyh@edgehill.ac.uk

Funding information
This work is supported by the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 20JG030) Special Construction Fund for Key Disciplines of Shaanxi Provincial Higher Education.

As a mental disorder, depression has affected people's lives, works, and so on. Researchers have proposed various industrial intelligent systems (IIS) in the pattern recognition field for audiovisual depression detection. This paper presents an end-to-end trainable intelligent system to generate high-level representations over the entire video clip. Specifically, a 3D-CNN equipped with a module Spatiotemporal Feature Aggregation Module (STFAM) is trained from scratch on AVEC2013 and AVEC2014 data, which can model the discriminative patterns closely related to depression. In the STFAM, channel and spatial attention mechanism and an aggregation method, namely 3D DEP-NetVLAD, are integrated to learn the compact characteristic based on the feature maps. Extensive experiments on the two databases (i.e., AVEC2013 and AVEC2014) are illustrated that the proposed intelligent system can efficiently model the underlying depression patterns and obtain better performances over the most video-based depression recognition approaches. Case studies are presented to describes the applicability of the proposed intelligent system for industrial intelligence.

*Equally contributing authors.
†Corresponding author.

KEYWORDS
Depression, Industrial intelligent system (IIS), 3D-CNN, Pattern recognition, Vector of local aggregated descriptors (VLAD)

1 | INTRODUCTION

With the speeding up the work and life, depression becomes a common mental disorder. According to the World Health Organization (WHO) report in 2020, depression has become a main contributor to the global mental disorder [1]. Mental health problems (i.e., depression, anxiety, etc.) have increased exponentially since the COVID-19 outbreak [2, 3, 4]. In some extreme cases, depression can bring about suicidal thoughts or attempts. So far, there are approximately 264 million people of all ages who suffer from depression[1]. Because depression can affect many aspects of our lives and works, there is an urgent requirement for estimating the severity of depression. Normally, depression manifestation is very complicated, which leads to difficulty in diagnosing. The current diagnostic approaches mainly depend on the self-reports of depressed patients and the judgments of the clinicians. They lack comprehensive and systematic strategies to incorporate behavioral observations that indicate psychological disorders. Moreover, the two ways mentioned above are subjective. Specifically, the evaluations made by clinicians can alter depending on their adopted diagnosis approaches [5], such as Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [6], the Beck Depression Inventory (BDI) [7], the 9–item Patient Health Questionnaire (PHQ–9) [8]. However, there is a conflict between the high prevalence of depression and the serious medical staff shortage. Therefore, it is essential to design automated industrial intelligent systems (IIS) to help clinicians promote efficiency.

To assist clinicians in effectively diagnose the severity of depression, researchers in the field of affective computing attempt to leverage the knowledge of artificial intelligence, psychology, physiology, and cognitive research to assess the severity of depression from audiovisual and physiological cues. According to the observation presented in [9], more than 50% patterns are around the facial region. Various systems have been proposed for estimating the severity of depression based on facial images [10, 11, 12, 13, 14, 15]. However, most of the above studies have proved to obtain superior performance in depression recognition tasks. However, there also exists some aspects for consideration. Firstly, many methods based on deep learning belong to frame (image) level [10, 11, 12, 13], while relatively rare studies consider video sequences as input [14]. The temporal patterns of videos are vital for representing the scale of depression. Secondly, frame (image) level-based methods mainly focus on extracting the deep spatial features closely related to depression. Recent works have adopted one-stream convolutional neural networks (CNN) to extract frame-level features and then utilize some simple pooling methods (i.e., average, max, etc.) to generate the video-level features [11]. However, frame-level features are not suitable for modeling the temporal patterns for depression recognition. Consequently, two-stream or multi-stream are proposed to model the spatial and temporal feature representations for further improving the performance of depression detection [14]. Though 3D convolutional neural network (3D-CNN) and recurrent neural network (RNN) technologies have been adopted to learn the spatiotemporal feature from facial images, and they adopt mean operation to pool the final features over the entire video. On the one hand, they do not consider the discriminative information of different sub-sequences, and the significant patterns may be lost during the mean operation performed. Thirdly, most previous works pre-train the deep models on large databases followed by fine-tuning with two depression databases (i.e., Audio/Visual Emotion Challenge (AVEC)2013, AVEC2014) [10, 11, 14], and the framework cannot be considered an end-to-end scheme for improving the depression efficiency.

[1] https://www.who.int/health-topics/depression

Interestingly, 3D-CNN technology has obtained great success in modeling the temporal patterns in video sequences [16, 17]. Moreover, Xu et al. propose to use 2D-CNN to learn the deep learned features and adopt vector locally aggregated descriptors (VLAD) to cluster the pattern of local features to obtain the global features across the entire video [18]. More recently, in [17], the authors propose to fine-tune the pre-trained 3D-CNN models to learn deep global representations for person re-identification [19, 20] (re-ID). In the works of [18, 17], a trainable VLAD layer is embedded in the architecture of the neural network (i.e.,2D-CNN, 3D-CNN) to learn more compact deep feature representations over frames or videos. Motivated by the successes of these works, we present a novel intelligent system for depression estimation, which includes the three integrated steps: first to extract the deep features by a 3D-CNN architecture, second to aggregate and encode the global representations via Depressed NetVLAD (DEP-NetVLAD) and attention mechanism, third to model high-level features over the entire video clip. Our goal is to consider the local feature extraction process compactly, aggregate the learned spatiotemporal features indiscriminately, and make the features a high-level representation for ensemble depression prediction. Towards the 3D-CNN architecture, which contains three cascade convolutional layers, to model local spatiotemporal patterns. The reason is that some subtle or discriminative patterns are not lost during the convolutional operations of 3D-CNN. For the spatiotemporal feature aggregation module (STFAM), which includes the attention mechanism, DEP-NetVLAD approaches encoding the mid-level feature representations. More importantly, DEP-NetVLAD is a trainable layer, which can be embedded and trained simultaneously along with the proposed end-to-end deep architecture. The spatial pyramid pooling (SPP) layer will build a high-level feature representation via transforming the multi-scale information from the local-global feature maps. A novel end-to-end depression scales prediction framework is proposed by closely integrating these three components. Mining both local and global characteristic information of depression is vital for a better depression recognition performance. Lastly, we conduct extensive experiments on the two public depression databases (i.e., AVEC2013 and AVEC2014) to validate the proposed method's effectiveness. The performances of the proposed IIS demonstrated that our study obtains comparable video-based depression recognition approaches.

1.1 | Contribution

The key contributions of this study can be outlined as follows:

1. Firstly, an end-to-end IIS is designed for effectively capturing the facial dynamics in the pattern recognition field as a non-verbal measurement for assessing the severity of depression.
2. Secondly, a 3D-CNN architecture is designed to extract the robust local spatiotemporal feature representations. Valuable features of pattern recognition for depression analysis are retained with the 3D-CNN.
3. Thirdly, a novel middle-level feature aggregation module STFAM is proposed. In the STFAM, channel attention and spatial attention are adopted for mining the salient patterns from the global and patch feature maps. More importantly, 3D DEP-NetVLAD is proposed to further aggregate the spatiotemporal features for ensemble depression recognition.
4. Fourth, extensive experiments using the AVEC2013 and AVEC2014 databases are performed to compare with other visual-based depression recognition methods. Results reveal that the proposed intelligent system is effective and more robust as compared to the state-of-the-art methods.
5. Finally, case studies are discussed to present the proposed intelligent system's applicability in industrial intelligence.

1.2 | Organization

The rest of this paper is organized as follows. Existing works on visual-based depression prediction are outlined in Section 2. Section 3 details the proposed end-to-end system. We introduce the databases and experimental results in Section 4. A case study in intelligent domains is introduced in Section 5. Conclusions and future works are introduced in Section 6.

2 | RELATED WORKS

Based on AVEC2013 and AVEC2014 sub-challenges, as well as the great success of deep learning technologies, various automatic depression recognition systems are designed. In the present paper, we focus on designing a new architecture based on visual cues for depression estimation. In the following, we first outline the visual features and introduce some feature aggregation methods for assessing the severity of depression.

2.1 | Hand-crafted vs Deep-learned Methods

This subsection briefly outlines hand-crafted and deep-learned features based on visual cues for depression recognition in previous works.

Generally speaking, hand-crafted methods mainly focus on extracting the hand-crafted features and using traditional machine learning models to predict the depression scale. For instance, for the hand-crafted features, local binary patterns from three orthogonal planes (LBP-TOP), etc. As a common feature descriptor, LBP-TOP has been used for facial expression [21], and depression recognition [22]. Several variants, for instance, the local gabor binary patterns from three orthogonal planes (LGBP-TOP) [23], the local phase quantisation from three orthogonal planes (LPQ-TOP) [24], median robust local binary patterns from three orthogonal planes (MRLBP-TOP) [15] and so on. Meanwhile, in the work of [12], the authors propose to extract low-dimensional descriptors as frame-level features based on human behavior primitives. In addition, two new feature descriptors based on spectral representations are proposed to indicate the multi-scale patterns of depression. For the traditional machine learning approaches used in the previous studies, e.g., decision trees (DT) [25], support vector regression (SVR) [26], partial least square (PLS) [27], etc.

Although the mentioned above approaches obtain excellent performance for depression recognition and analysis. However, there are still some limitations of hand-crafted methods for predicting the level of depression severity. Firstly, to develop hand-crated features needs a certain accumulation of domain knowledge. For instance, LBP-TOP has been valid to obtain good performances in many tasks, such as facial expression, depression recognition [22], etc. Secondly, hand-crafted features may ignore certain patterns about depression assessment. In other words, the significant characteristic can not be mined well in the audiovisual signals. Furthermore, the idea of the developed features depends on researchers' subjective perspectives.

In recent years, deep learning technology has been successfully and widely adopted in various communities. For depression recognition, various studies have been used deep learning to learn the deep representations from visual cues [10, 28, 14, 11, 13, 12].

Zhu et al. [10] train a two steam CNN architecture for combining facial appearance and dynamics to assess the scale of depression recognition. They reported better results than other visual-based methods.

In [11], the authors propose DepressNet, a deep regression network for predicting the severity of depression from single images. A Depression activation map (DAM) is also adopted to represent important facial image regions to determine the scale of depression. Meanwhile, they also design a multi-region DepressNet to model the different

regions' different patterns to improve overall results.

In [12], the authors extract a new feature descriptor from each frame and propose spectral heatmaps and spectral vectors to model the discriminative representations based on action units (AU). The spectral representations are input into the convolution neural networks (CNN) and artificial neural networks (ANN) for predicting the depression scale. Extensive experiments are carried out on the two depression databases (i.e., AVEC2013 and AVEC2014), and they achieved superior performance in the depression recognition task.

In [13], a two-stream framework is designed to model the spatiotemporal representations for depression recognition. The temporal median pooling (TMP) method is adopted to model some temporal patterns of the generated features via CNN. Lastly, experimental results of the two depression databases (i.e., AVEC2013 and AVEC2014) shown the proposed method's efficiency.

In [14], they extract local-global features of convolutional 3D networks to improve the overall performance. The proposed network is equipped with 3D global average pooling to represent spatiotemporal features for depression detection. Experimental results show that fusing the local and global features of C3D networks obtains promising performance.

In [28], an AI system is designed to estimate the scale of depression. The system can fuse the complementary patterns between the hand-crafted and deep-learned features. For visual cues, deep-learned features are extracted, which include some discriminative information related to depression. Temporal movements on the different feature spaces are described by feature dynamic history histogram (FDHH).

Most of the methods always fine-tuned the deep models for deep-learned features, which have been pre-trained on a large database (e.g., CASIA WebFace Database, etc.). However, the existing methods have not been considered as an end-to-end scheme for depression recognition. Our goal is to explore an IIS in this study, which can directly predict the depression scale from facial image sequences. More importantly, our proposed framework is an end-to-end system to predict the severity level of depression directly.

2.2 | Feature Aggregation Methods

As mentioned in the reference [15], for hand-crafted features, various feature aggregations are proposed to aggregate the frame-level audiovisual features, such as a bag of words (BoW), VLAD, and fisher vector (FV), etc.

Meanwhile, as mentioned in Section 2.1, deep learning has been proven to obtain excellent performance in many communities. Therefore, in [29], the authors design a novel sequential vector of locally aggregated descriptor (VLAD) layer, denoted as SeqVLAD, combine with Recurrent Convolution Networks (RCNs) to generate an ensemble framework for video-based action recognition. In [30], spatial pyramid-enhanced VLAD (SPE-VLAD) method and weighted T-loss (WT-loss) function are combined with VGG-16 network to constitute an end-to-end architecture for place recognition. In [31], action-stage (ActionS) emphasized spatiotemporal vector of locally aggregated descriptors (ActionS-ST-VLAD) approach is proposed to aggregate the global features over entire video for action recognition. Inspired by these works, we propose using NetVLAD to model the deep global feature representations for depression recognition. As far as we know, this is the first work to adopt NetVLAD to represent the discriminate patterns related to depression.

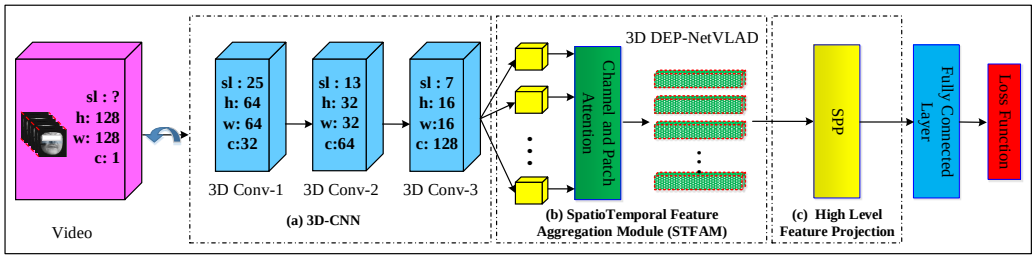

FIGURE 1 The pipeline of the proposed IIS for the diagnosis of depression. (a) The local spatiotemporal features are extracted by 3D-CNN blocks. (b) Feature aggregation operation is performed on the local spatiotemporal features by the STFAM. (c) High-level features are learned by SPP.

3 | OUR PROPOSED APPROACH

The proposed approach for predicting the scale of depression is illustrated in Fig. 2. The facial region is cropped and aligned by the OpenFace toolbox. To describe the proposed framework distinctly, we separate the proposed framework into four steps. Firstly, the local spatiotemporal features are extracted. Secondly, to obtain a discriminative and intermediate representation, a novel module STFAM is proposed. In the STFAM, DEP-NetVLAD is proposed to aggregate the informative and global patterns related closely to depression. Thirdly, to make a high-level representation, SPP is adopted for the final prediction. Of particular importance to the proposed method is an end-to-end scheme for depression recognition. In what follows, local deep spatiotemporal features are extracted by 3D-CNN for intermediate feature representation in Section 3.1; then, we present the introduced STFAM in Section 3.2; Lastly, SPP is described in Section 3.3.

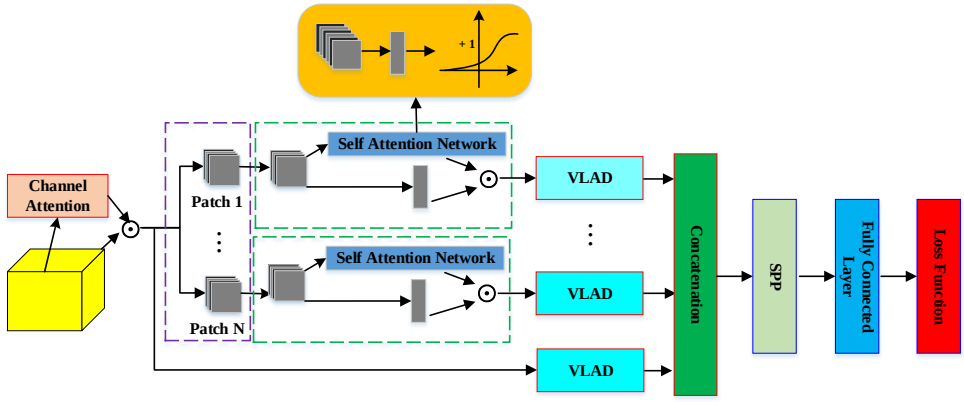

FIGURE 2 The detailed illustration for the STFAM.

3.1 | Learning Local Spatiotemporal Feature Representations

In the presented paper, we assume that global feature representations can be modeled by local spatiotemporal feature representations. 3D-CNN has been confirmed to obtain promising performance for modeling sequences issues [16, 17]. Therefore, we develop a 3D-CNN architecture to mine discriminative spatiotemporal features. In the 3D-CNN architecture, we only adopt three 3D convolutional operations, capturing both the spatial and temporal patterns related to the depression. As illustrated in Fig.1, it can be seen that the video is first pre-processed by the OpenFace toolkit with the size of 128×128. In our work, the gray channel is adopted to extract the image features. The length of the video sequence is alterable from one to another (represented as "?" in Fig. 1). For the traditional 3D-CNN, the length of the input is 16 frames. In our task, to learn the long-term information of depression, we consider 25 frames/s as a mini-batch sub-sequence to perform over the entire video. Hence, if the length of the video is 1000 frames, we need to loop 40 times to process the entire video clip for the feature extraction step.

The 3D-CNN architecture consists of three successive convolutional layers. For the first layer, the linear activation function is considered, and the filter size is $5 \times 5 \times 5$, where the first dimension is the temporal information, while the second and third dimension denotes the spatial information. And the stride of this layer is set to $2 \times 2 \times 2$ to perform the spatiotemporal dimensions. The filters of size are set to 30. Normalization operation is performed to normalize the activities of the neurons. After that, a dropout operation is also adopted to prevent overfitting and to heighten generalization. For the second layer, the filter size of $3 \times 3 \times 3$ is designed, and the size of the feature map is set to 60. Additionally, we consider ReLU as an activation function. Also, normalization and dropout operations are also performed at this layer. For the third layer, we perform the same configuration as the second layer. Only the difference is that the number of filters is increased to 90.

In our task, the input is initialized at time-step t_1 with size of $25 \times 128 \times 128 \times 1$, after pass the 3D-CNN block, the output is $4 \times 16 \times 16 \times 90$. By using the 3D-CNN block, deep local spatiotemporal features are extracted to obtain a compact representation for estimating the severity level of depression.

3.2 | Aggregating Mid-level Spatiotemporal Representations

To learn compact and discriminative representation from the local spatiotemporal features extracted by the previous step, we divide the STFAM into two parts: the first part is channel and patch attention, and the second part is DEP-NetVLAD.

The attention mechanism is proposed to represent the salient patterns in many raw signals (e.g., audio, video, text, etc.). To retain the valuable characteristic and filter the feature maps' redundant information (the blue cube in Fig. 2), channel attention is performed. For a convenient description, we assume that a four-dim tensor $\mathbf{Fe}(l, h, w, c)$ be the Conv-3 feature cubic, where t represents the temporal dimension, h and w are the height, weight of feature map, c is number of channels. The channel attention learns a three-dimension weights $W_t(h, w, c)$ at the time step t. Thus, the weights $W_t(h, w, c)$ is performed on the feature $\mathbf{Fe}(l, h, w, c)$ to generate the new discriminative features:

$$\mathbf{Z}_{t_{global}} = \mathbf{Fe} \cdot W_t \tag{1}$$

Then, to capture the salient representations of the features map **Fe**, we divide the feature maps at the spatial direction into many patches to automatically learn the characteristic. In our task, we use the different size of patches, i.e., 2×2, 4×4, 8×8. Therefore, if the patch size is 2×2, the cubic feature with the size of $[16 \times 16 \times 128]$ is divided into 8 patches, and so on. After that, the feature maps are input into two branches. The first branch considers the feature

maps as the vector level local feature. For the second branch, a self-attention net is proposed to concentrate on the spatial patches' discriminative representation regions. The self-attention net is consists of one convolution operation, one fully connected layer, and a sigmoid operation. The sigmoid function is adopted to restrict the range of the output ranges between 0 and 1. The learned weight α_j can be defined as:

$$\alpha_t = \omega_j(Z_{t_{patch}}) \tag{2}$$

where α_t is a scalar, ω_j represent the operations of self attention net.

Then, we perform α_t on the patch feature map $\mathbf{Z}_{t_{patch_j}}$ to generate the discriminative feature:

$$\mathbf{Z}_{t_{patch_j}} = \alpha_t \cdot \mathbf{Z}_{t_{patch_j}} \tag{3}$$

The second step is to aggregate the discriminative features, which obtained from the global and patch feature maps. As illustrated by blue box in Fig. 2, the VLAD aggregation method is adopted. Let assume that the feature cubic $\mathbf{Z}_{t_{global}} \in \mathbb{R}^D, \mathbf{Z}_{t_{patch_j}} \in \mathbb{R}^D$ to be denoted as an anchor point $\{\mathbf{c}_k\}, k \in K$. The feature space $\mathbb{R}^D$ can be divided into K cells according to K depression words. The VLAD aggregation method stores the sum of residuals (difference vector between the input feature $\mathbf{Z}_{t_{global}} \in \mathbb{R}^D, \mathbf{Z}_{t_{patch_j}} \in \mathbb{R}^D$ and its corresponding cluster center $\mathbf{c}_k$) for each visual word $\mathbf{c}_k$. Note that we only use $\mathbf{Z}_{t_{patch_j}} \in \mathbb{R}^D$ to describe the proposed method.

In our task, we aim to design a trainable VLAD layer that plug into 3D-CNN to form an end-to-end tool for depression assessment. Following the work [17], we also adopt the soft assignment strategy to multiple clusters, i.e., $\alpha_k(\mathbf{Z}_{t_{patch_j}}) = (e^{-\alpha||\mathbf{Z}_{t_{patch_j}} - c_k||^2})/\sum_{k'} e^{-\alpha||\mathbf{Z}_{t_{patch_j}} - c_{k'}||^2}$, which assigns the weight of $\mathbf{Z}_{t_{patch_j}}$ to the cluster $\mathbf{c}_k$ proportional to their proximity. α is a hyperparameter that can be tuned. Therefore, the feature maps $\mathbf{Z}_{t_{patch_j}}$ can be represented by $\mathbf{Z}_{t_{patch_j}} - \mathbf{c}_k$. The representation of entire video can be defined by:

$$\mathbf{V}[j,k] = \underbrace{\sum_{n=1}^{N+1} \frac{(e^{-\alpha||\mathbf{Z}_{t_{patch_j}} - c_k||^2})}{\sum_{k'} e^{-\alpha||\mathbf{Z}_{t_{patch_j}} - c_{k'}||^2}}}_{soft-assignment} \underbrace{(\mathbf{Z}_{t_{patch_j}} - c_k)}_{Residual} \tag{4}$$

In Equ. 4, the left term is the soft assignment of $\mathbf{Z}_{t_{patch_j}} - \mathbf{c}_k$ to cell k, and the right term is the residual between the feature map and the anchor. By performing the sum operation, the final aggregated representation with the vector size of KD. Finally, we adopt a L2-normalization to $\mathbf{V}[j,k]$. Thus, the proposed 3D DEP-NetVLAD is developed to aggregate global and patch feature representations into a discriminative and compact pattern over the entire video clip.

3.3 | Learning High-level Features

To obtain the scale-invariant and high-level representation, a SPP layer is adopted to represent multi-scales on top of the output of STFAM. The idea of SPP is to segment the feature map into different divisions from finer to coarser scales, finalized with an aggregation of local features. The SPP layer also can promote scale-invariance and reduce the issue of overfitting. The finer resolutions' generated features are associated with a heavier weight and coarser

resolutions with a lower weight. Specifically, the weight of the multi-scale features can be written as:

$$\frac{1}{2^S} M^0 + \sum_{S=1}^{S} \frac{1}{2^{S-l+1}} M^l \tag{5}$$

where S represents the number of levels, $S = 1$ represents the current level. M^0 denotes the feature map at coarse resolution, and M^S represents the feature map at fine resolution.

In every spatial bin, we adopt max-pooling to aggregate the responses of each filter. The output of SPP is the sum of the total number of bins. In our task, the shape of the output of STFAM is [batch_size, 2048], then we obtain a weighted feature vector of size 150-d. The window sizes of SPP is 80×1, 40×1, 20×1, 10×1 and their corresponding stride is 80×1, 40×1, 20×1, 10×1, respectively. The fixed-dimensional vectors are then inputted into the fully-connected layer.

As a deep network, the loss function plays an important part in the final classification or regression. In our task, depression detection can be regarded as a regression problem of machine learning. The root mean square error (RMSE) and the mean absolute error (MAE) is adopted to measure the performance of depression assessment. Hence, we use RMSE instead of MSE as the loss function, which is considered suitable for our work. The RMSE loss L_{dep} is give by:

$$L_{dep} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (p_j - \tilde{p}_j)^2} \tag{6}$$

where N, $\hat{p_j}$ and p_j represent the number of samples, the final BDI prediction, and the ground truth, respectively.

4 | EXPERIMENTS

We perform extensive experiments to evaluate our proposed architecture. In Section 4.1, the evaluated databases (i.e., AVEC2013 and AVEC2014) are briefly introduced. In Section 4.2, the experimental setup and evaluation measures are also shown. Lastly, we show the experimental results and provide some discussion in Section 4.3.

4.1 | Databases

In the present paper, all experiments are validated on two publicly depression databases, i.e., AVEC2013 and AVEC2014. The average age of the participant is 31.5 years (age ranging from 18-63). A webcam and a microphone are used for recording the audio and appearance signals. BDI-II is used as labeling for each sample.

In the AVEC2013 depression corpus, there are a total of 150 video clips from 82 subjects. The audiovisual samples have been divided into three partitions by the publisher, i.e., training, development, and test set. For every partition, it has 50 recordings.

For the AVEC2014 depression corpus, there are two tasks included, i.e., Freeform and Northwind. For the two tasks, there are 150 video clips from 84 subjects. The same as AVEC2013, it also has three partitions, i.e., training, development, and test sets. Therefore, there are 100 samples in the partitions.

4.2 | Experimental Setup and Evaluation Measures

4.2.1 | Experimental Setup

Both on the two depression databases, face detection, and landmarks localization operation are performed. Then, the aligned facial images are resized and cropped of the size 128×128. To guarantee the accuracy of depression estimation, we check each frame manually. Specifically, if the facial region is not detected in the image frame, the corresponding frame is filtered. For AVEC2013, the faces are found only of 95.61% of the frames, while the rate with 99.14% for the AVEC2014 database.

For the visual words of NetVLAD, we consider various numbers of 8, 16, 32, and 64. For the size of patch, we adopt different size of 2×2, 4×4, and 8×8. Therefore, we obtain 12 combinations for each video clip.

TensorFlow deep learning platform is used to train the proposed framework. To get fast convergence, the adaptive moment estimation (ADAM) optimizer is adopted to train our proposed architecture, which has been considered as an essential method to represent and evaluate deep architecture on small databases. The learning rate is set to 0.000001. The batch size is set to 32. We conduct the experiments with two Titan-X GPU (each with 12G memory). The number of iterations is empirically set to 60k.

4.2.2 | Evaluation Measures

To make a fair comparison, the mean absolute error (MAE) and root mean square error (RMSE) are adopted for measuring the performance of depression recognition methods, as shown in Equ.7 and Equ. 8, where N, p_i and $\tilde{p}_i$ represent the number of samples, the ground truth, and the predicted value of the j-th subjects, respectively.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|p_j - \tilde{p}_j\right| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(p_j - \tilde{p}_j)^2} \tag{8}$$

4.3 | Experimental Results

Different combinations of the proposed framework are designed to evaluate each part of the introduced framework's availability. Meanwhile, we compare NetVLAD with other aggregation methods, i.e., NetFV, NetRVLAD, and NetSoftBow, to show the proposed architecture's efficiency. Lastly, the proposed framework is further compared with the most video-based depression recognition approaches to illustrate its excellent performance.

4.3.1 | Performance of Depression Recognition Using Different Aggregations Approaches

This section adopts different parameters to evaluate the proposed method and compare it with other feature aggregation approaches for depression detection. As illustrated in Tables 1 and 2, the different combinations of parameters are performed on the proposed framework on the two databases (i.e., AVEC2013 and AVEC2014), respectively. From

TABLE 1 Performance of different combinations of the proposed framework on the test set of AVEC2013.

Patch	Depression Words	RMSE	MAE
2×2	8	9.31	7.80
2×2	16	9.57	7.85
2×2	32	8.91	7.12
2×2	64	9.09	7.88
4×4	8	8.74	7.28
4×4	16	**8.46**	**6.83**
4×4	32	8.47	6.85
4×4	64	9.35	7.69
8×8	8	9.24	7.89
8×8	16	9.06	7.60
8×8	32	9.33	7.58
8×8	64	9.22	7.59

TABLE 2 Performance of different combinations of the proposed framework on the test set of AVEC2014.

Patch	Depression Words	RMSE	MAE
2×2	8	9.19	7.69
2×2	16	8.66	7.42
2×2	32	8.96	7.22
2×2	64	9.02	7.46
4×4	8	9.00	7.56
4×4	16	**8.42**	**6.78**
4×4	32	8.49	7.12
4×4	64	8.98	7.44
8×8	8	9.17	7.47
8×8	16	9.06	7.57
8×8	32	9.13	7.59
8×8	64	9.08	7.32

table 1, we can notice that, on the AVEC2013 database, the patch size of 4×4 and depression word of 16 obtain the best depression performance, with the RMSE of 8.46, MAE of 6.83.

While on the AVEC2014 database, we also perform various experiments to verify the proposed scheme's capability further. From Table 2, one can notice that the patch size of 4×4 and depression word of 16 also get the best depression recognition results, with the RMSE of 8.42 and MAE of 6.78. From the performance, we can observe

TABLE 3 Comparison with another aggregation methods for depression estimation on AVEC2013.

	Patch Size	Visual Words	RMSE	MAE
NetVLAD	4×4	16	**8.46**	**6.83**
NetRVLAD	4×4	16	10.12	8.41
NetFV	4×4	16	9.78	8.40
NetSoftBow	4×4	16	9.53	8.32

TABLE 4 Comparison with another aggregation methods for depression estimation on AVEC2014.

	Patch Size	Visual Words	RMSE	MAE
NetVLAD	4×4	16	**8.42**	**6.78**
NetRVLAD	4×4	16	9.91	8.29
NetFV	4×4	16	9.71	8.17
NetSoftBow	4×4	16	9.47	8.07

that, on the one hand, the combination of the parameters is efficient for our task; on the other hand, the results have further validated the capability of the proposed architecture for estimating the severity of depression. The performances on the two depression databases show that the efficiency of the proposed method to assess the severity scale of depression from video sequences.

Afterward, to illustrate the efficiency of the DEP-NetVLAD, we replace NetVLAD with NetFV, NetRVLAD, and NetSoftBow, and keep other parts fixed of the framework. It is noticed that the optimal parameters are adopted from NetVLAD to compare the performance among other aggregation methods. As presented in Tables 3 and 4, one can observe that, both on the two depression databases (i.e., AVEC2013 and AVEC2014), the NetVLAD obtains the best depression recognition performance.

4.3.2 | Comparison With State-Of-The-Arts

In this part, we compare our proposed framework with other previous approaches on the two depression databases in Table 5 and 6, respectively. On the AVEC2013 database, as shown in Table 5, our method yields comparable performances that the most of the video-based results, except for the approach of [35], [13], [11]. In [35], the authors propose adopting convolutional 3D networks that are pre-trained on the CASIA dataset to represent spatio-temporal feature representations, where deep architecture is adopted for estimating the severity of depression. In [13], the authors also fine-tune the pre-trained models to assess the severity of depression. In [11], a deep depression recognition architecture is proposed to model the discriminate features from facial images. By comparing the works of Zhou et al. [11], Melo et al. [35](C3D) and Melo et al. [36](ResNet-50), the RMSEs are not surpass of them. Our method is trained from scratch that is an end-to-end scheme for depression recognition. Meanwhile, our method does not have to be leveraged the pre-trained models for depression recognition. In particular, the works of Zhou et al. [11] and Melo et al. [36] (ResNet-50), the authors fine-tuned the large pre-trained deep models on AVEC2013 and AVEC2014 databases from facial images for depression recognition. The possible reason is that the large deep models not only contain typical features to model the severity of depression, but also reduce the cost of re-training the deep archi-

TABLE 5 Performance of different architecture for visual-based automatic depression diagnosis on the test set of AVEC2013.

Methods	RMSE	MAE
Baseline [24]/ LPQ, SVR	13.61	10.88
Cummins et al. [32]/ STIP and PHOG, SVR	10.45	N/A
Meng et al. [33]/ EOH and LBP, PLSR	11.19	9.14
Wen et al. [34]/ LPQ-TOP, SVR	10.27	8.22
Zhu et al. [10]/ Optical Flow, 2D-CNN	9.82	7.58
Mohamad et al.[14]/ C3D, RNN	9.28	7.37
He et al. [15]/ MRLBP-TOP, DPFV, SVR	9.20	7.55
Zhou et al. [11]/ 2D-CNN	8.19	6.30
Melo et al. [35]/ C3D	8.26	6.40
Melo et al. [36]/ ResNet-50	8.25	6.30
Md et al. [13]/ LSTM	8.93	7.04
Proposed Approach /3D-CNN, STFAM, SPP	8.46	6.83

TABLE 6 Performance of different architecture for visual-based automatic depression diagnosis on the test set of AVEC2014.

Methods	RMSE	MAE
Baseline [23]/ LGBP-TOP, SVR	10.86	8.86
Jan et al. [27]/ EOH, LBP and LPQ, PLSR	10.50	8.44
Zhu et al. [10]/ Optical Flow, 2D-CNN	9.55	7.47
Mohamad et al.[14]/ C3D, RNN	9.20	7.22
He et al. [15]/ MRLBP-TOP, DPFV, SVR	9.01	7.21
Zhou et al. [11]/ 2D-CNN	8.39	6.21
Melo et al. [35]/ C3D	8.31	6.59
Melo et al. [36]/ ResNet-50	8.23	6.13
Md et al. [13]/ LSTM	8.78	6.86
Proposed Approach /3D-CNN, STFAM, SPP	8.42	6.78

tecture for depression recognition. However, our method has the following three advantages: 1) Temporal patterns of depression is modeled by 3D-CNN deep learning technology; 2) Valuable and discriminative features helpful for depression analysis are retained and aggregated with the Net-VLAD; 3) The most important is that an end-to-end architecture is proposed to assist the clinicians to assess the severity of depression.

Similar to AVEC2013, on the AVEC2014 database, as shown in Table 6, our method obtains comparable results to most of the video-based depression recognition methods, with 6.78 MAE and 8.42 RMSE on the test set. Based

on the experimental performances, we can draw the conclusion that the proposed method can automatically learn local and global characteristic information of the facial region and outperform most of the state-of-the-art methods for depression recognition on AVEC2013 and AVEC2014 databases. This observation further illustrates the effectiveness of the proposed scheme for predicting the scale of depression severity.

5 | CASE STUDY IN INDUSTRIAL INTELLIGENCE

The majority of industrial intelligence use-cases for detecting depression can be grouped into four major categories, namely, virtual counseling, patient monitoring, precision therapy, and self-assessment. These categories are discussed as follows.

1) Virtual Counseling: AI tools are designed for depression recognition, such as Woebot is developed to help depressed patients adjust their mood and the representation of depression. Another variation Wysa has been adopted by more than 200,000 users from 30 countries. Besides, the way of virtual counseling is a benefit of the clinical application of clinicians. Moreover, virtual counseling is vital for depression recognition and analysis.

2) Patient Monitoring: Ginger.io is founded to adopt machine learning and pattern recognition methods to support the users. Besides, Sunrise Health develops an APP to manage the patient's activity and provide a warning for them. Furthermore, patient monitoring research has significant importance in the clinic. In particular, the symptoms are vital for diagnosing the severity of depression in the early stage. In our case, the developed architecture can assist clinicians in assessing the severity of depression.

3) Precision Therapy: Mindstrong Health was founded to use machine learning and pattern recognition technologies to assist and assess mood disorders by collecting data from smartphone devices. As mentioned in 1) and 2), the available tools are significant for depression estimation in clinical application.

4) Self-assessment: In some cases, depressed subjects need to assess their symptoms by themselves. Especially the high prevalence of depression now, an appropriate self-assessment methodology is important for estimating the severity of depression. Hence, the proposed architecture can assess the severity of depression and model the discriminate features to help clinicians diagnose the severity of depression.

The discussion as mentioned above reveals that some tools are available which had been used successfully for counseling, monitoring, and therapy. Further, this discussion describes the necessity of developing a novel IIS based on facial regions for depression recognition and diagnosis in industrial applications.

In particular, researchers are increasingly developing social media networks' potential as tools to estimate depression and assess its symptoms as manifested in user comments and related activities (audio, video, etc.). Social networks such as Twitter.Inc, Facebook.Inc, and Reddit.Inc has become part of our daily lives as media through which to share our thoughts, feelings, and overall emotional status. Hence, in these contexts, our proposed IIS based on pattern recognition methods will be used by individuals and hospitals in the future.

6 | CONCLUSIONS AND FUTURE WORKS

With the speeding up of work and life, depression becomes a common mental disorder. Peoples have been suffering from depression disorders. Moreover, after out-breaking the COVID-19 pandemic in 2020, more individuals have specific symptoms of depression. This paper proposed an end-to-end IIS based on pattern recognition to generate high-level features over the entire video clip for video-based depression recognition. We implemented a 3D-CNN block along with a trainable module, STFAM, and SPP to "encode" the spatiotemporal patterns around the facial

regions into a high-level representation. Specifically, a 3D-CNN equipped with a module STFAM was trained from scratch on AVEC2013 and AVEC2014 data, closely modeling the discriminative patterns related to depression. In the STFAM, channel and spatial attention mechanism and an aggregation method, namely 3D DEP-NetVLAD, were integrated to learn the compact characteristic based on the feature maps. Finally, we validated the proposed system on the two depression databases (i.e., AVEC2013 and AVEC2014), and obtained promising depression estimation performances. And the introduced IIS can be adopted in industrial scenarios, such as hospitals, psychiatric clinics, schools, etc. When fusing multiple modalities, it is significant to classify each modality's contribution in the depression prediction, and attention networks can be adopted to learn the relative importance of depression patterns. The same as the work of [37], we will focus on multimodal fusion (audio, video, text, etc.) scheme based on pattern recognition for depression recognition and analysis in the future. Besides, we will use Transformer [38] to model the discriminative patterns for depression recognition. Moreover, we try to adopt the IIS in industrial domains.

Conflict of Interest

The authors declare that there are no conflict of interests.

AUTHOR CONTRIBUTIONS

All authors contributed to the preparation of this manuscript. Lang He, Chenguang Guo, and Prayag Tiwari share equal contributions as first co-authors. Lang He, Chenguang Guo, and Prayag Tiwari contributed to the methodology, experiments, and writing manuscript. Hari Mohan Pandey and Wei Dang assisted in methodology, writing the manuscript, and proofreading.

ORCID

Lang He (https://orcid.org/0000-0003-2515-8579)
Chenguang Guo (https://orcid.org/0000-0002-5711-7977)
Prayag Tiwari (https://orcid.org/0000-0002-2851-4260)
Hari Mohan Pandey (https://orcid.org/0000-0002-9128-068X)
Wei Dang (https://orcid.org/0000-0003-2477-8485)

references

[1] Mathers C, Fat DM, Boerma JT. Depression and Other Common Mental Disorders:Global Health Estimates. World Health Organization; 2020.

[2] Ashraf S, Abdullah S. Emergency decision support modeling for COVID-19 based on spherical fuzzy information. International Journal of Intelligent Systems 2020;35(11):1601–1645.

[3] Chai J, Xian S, Lu S. Z-uncertain probabilistic linguistic variables and its application in emergency decision making for treatment of COVID-19 patients. International Journal of Intelligent Systems 2021;36(1):362–402.

[4] Gribova V, Shalfeeva E. Ontology of anomalous processes diagnosis. International Journal of Intelligent Systems 2021;36(1):291–312.

[5] Tiwari P, Uprety S, Dehdashti S, Hossain MS. TermInformer: unsupervised term mining and analysis in biomedical literature. Neural Computing and Applications 2020;p. 1–14.

[6] Bogduk N. Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association,; 2013.

[7] Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. Journal of Personality Assessment 1996;67(3):588–597.

[8] Kroenke K, Spitzer RL. The PHQ-9: A New Depression Diagnostic and Severity Measure. Psychiatric Annals 2002;32(9):509–515.

[9] Ellgring H. Non-verbal Communication in Depression. Cambridge University Press; 2007.

[10] Zhu Y, Shang Y, Shao Z, Guo G. Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics. IEEE Transactions on Affective Computing 2017;9(4):578–584.

[11] Zhou X, Jin K, Shang Y, Guo G. Visually Interpretable Representation Learning for Depression Recognition from Facial Images. IEEE Transactions on Affective Computing 2018;p. 1–1.

[12] Song S, Jaiswal S, Shen L, Valstar M. Spectral Representation of Behaviour Primitives for Depression Analysis. IEEE Transactions on Affective Computing 2020;p. 1–1.

[13] Depression Level Prediction Using Deep Spatiotemporal Features and Multilayer Bi-LTSM, author=Md Azher Uddin, Joolekha Bibi Joolee, and Young-Koo Lee. IEEE Transactions on Affective Computing 2020;p. 1–1.

[14] Al Jazaery M, Guo G. Video-based Depression Level Analysis by Encoding Deep Spatiotemporal Features. IEEE Transactions on Affective Computing 2018;p. 1–1.

[15] He L, Jiang D, Sahli H. Automatic Depression Analysis Using Dynamic Facial Appearance Descriptor and Dirichlet Process Fisher Encoding. IEEE Transactions on Multimedia 2019;21(6):1476–1486.

[16] Ji S, Xu W, Yang M, Yu K. 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 2013;35(1):221–231.

[17] Wu L, Wang Y, Shao L, Wang M. 3-D PersonVLAD: Learning Deep Global Representations for Video-Based Person Reidentification. IEEE Transactions on Neural Networks and Learning Systems 2019;30(11):3347–3359.

[18] Xu Z, Yang Y, Hauptmann AG. A Discriminative CNN Video Representation for Event Detection; 2015. p. 1798–1807.

[19] Wu J, Yang Y, Lei Z, Wang J, Li SZ, Pandey HM. An end-to-end exemplar association for unsupervised person Re-identification. Neural Networks 2020;129:43–54.

[20] Yang Y, Zhang T, Cheng J, Hou Z, Tiwari P, Pandey HM, et al. Cross-modality paired-images generation and augmentation for RGB-infrared person re-identification. Neural Networks 2020;128:294–304.

[21] Zhao G, Pietikainen M. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 2007;29(6):915–928.

[22] Dhall A, Goecke R. A Temporally Piece-wise Fisher Vector Approach for Depression Analysis. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII); 2015. p. 255–259.

[23] Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM; 2014. p. 3–10.

[24] Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, et al. AVEC 2013: The Continuous Audio/visual Emotion and Depression Recognition Challenge. In: Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge ACM; 2013. p. 3–10.

[25] Yang L, Jiang D, Han W, Sahli H. DCNN and DNN based Multi-modal Depression Recognition; 2017. p. 484–489.

[26] He L, Jiang D, Sahli H. Multimodal Depression Recognition With Dynamic Visual and Audio cues; 2015. p. 260–266.

[27] Jan A, Meng H, Gaus YFA, Zhang F, Turabzadeh S. Automatic Depression Scale Prediction using Facial Expression Dynamics and Regression. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge ACM; 2014. p. 73–80.

[28] Jan A, Meng H, Gaus YFBA, Zhang F. Artificial Intelligent System for Automatic Depression Level Analysis Through Visual and Vocal Expressions. IEEE Transactions on Cognitive and Developmental Systems 2017;10(3):668–680.

[29] Xu Y, Han Y, Hong R, Tian Q. Sequential Video VLAD: Training the Aggregation Locally and Temporally. IEEE Transactions on Image Processing 2018;27(10):4933–4944.

[30] Yu J, Zhu C, Zhang J, Huang Q, Tao D. Spatial Pyramid-Enhanced NetVLAD With Weighted Triplet Loss for Place Recognition. IEEE Transactions on Neural Networks and Learning Systems 2020;31(2):661–674.

[31] Tu Z, Li H, Zhang D, Dauwels J, Li B, Yuan J. Action-Stage Emphasized Spatiotemporal VLAD for Video Action Recognition. IEEE Transactions on Image Processing 2019;28(6):2799–2812.

[32] Cummins N, Joshi J, Dhall A, Sethu V, Goecke R, Epps J. Diagnosis of Depression by Behavioural Signals: A Multimodal Approach. In: Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge ACM, Barcelona, Spain: ACM; 2013. p. 11–20.

[33] Meng H, Huang D, Wang H, Yang H, AI-Shuraifi M, Wang Y. Depression Recognition Based on Dynamic Facial and Vocal Expression Features using Partial Least Square Regression. In: Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge ACM, Barcelona, Spain: ACM; 2013. p. 21–30.

[34] Wen L, Li X, Guo G, Zhu Y. Automated Depression Diagnosis Based on Facial Dynamic Analysis and Sparse Coding. IEEE Transactions on Information Forensics and Security 2015;10(7):1432–1441.

[35] de Melo WC, Granger E, Hadid A. Combining Global and Local Convolutional 3D networks for Detecting Depression from Facial Expressions. FG; 2019. p. 1–8.

[36] de Melo WC, Granger E, Hadid A. Depression Detection Based on Deep Distribution Learning. In: 2019 IEEE International Conference on Image Processing (ICIP); 2019. p. 4544–4548.

[37] Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multi-level attention network using text, audio and video for depression prediction. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; 2019. p. 81–88.

[38] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017. p. 6000–6010.