Version of Record: https://www.sciencedirect.com/science/article/pii/S0925231220315101
Manuscript_6f5df7230f323e1989042623c9af270b

Automatic Depression Recognition using CNN with Attention Mechanism from Videos

Lang He[a,b], Jonathan Cheung-Wai Chan[c], Zhongmin Wang[a,b]

[a] *School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi 710121, China*
[b] *Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an Shaanxi 710121, China*
[c] *Department of Electronics and Informatics, Vrije Universiteit Brussel(VUB), 1050 Brussel, Belgium*

Abstract

Artificial intelligence (AI) has incorporated various automatic systems and frameworks to diagnose the severity of depression using hand-crafted features. However, process of feature selection needs domain knowledge and is still time-consuming and subjective. Deep learning technology has been successfully adopted for depression recognition. Most previous works pre-train the deep models on large databases followed by fine-tuning with depression databases (i.e., AVEC2013, AVEC2014). In the present paper we propose an integrated framework - Deep Local Global Attention Convolutional Neural Network (DLGA-CNN) for depression recognition, which adopts CNN with attention mechanism as well as weighted spatial pyramid pooling (WSPP) to learn a deep and global representation. Two branches are introduced: Local Attention based CNN (LA-CNN) focuses on the local patches, while Global Attention based CNN (GA-CNN) learns the global patterns from the entire facial region. To capture the complementary information between the two branches, Local-Global Attention-based CNN (LGA-CNN) is proposed. After feature aggregation, WSPP is used to learn the depression patterns. Comprehensive experiments on AVEC2013 and AVEC2014 depression databases have demonstrated that the proposed method is capable of mining the underlying depression patterns of facial videos and

Email address: langhe@xupt.edu.cn (Lang He)

outperforms the most of the state-of-the-art video-based depression recognition approaches.

Keywords: Depression, CNN with attention mechanism, Local Attention based CNN (LA-CNN), Global Attention based CNN (GA-CNN)

1. Introduction

Depression ranks the fourth among these most serious mental health issues by 2020 [1]. In general cases, it not only has mild harm to individual life, but also has a certain influence for family and society. In some worst cases, it may lead to suicide. Therefore, it is urgent to find an efficient solution to diagnose and treat the clinical depression.

In recent years, many methods from various perspectives to assist psychologists or clinicians to diagnose and treat the clinical depression have been developed, mainly from affective computing, computer vision and machine learning communities and so on. To predict the severity of depression based on audio-visual cues, traditional approaches commonly consist of three successive procedures: 1) feature extraction, 2) feature aggregation, and 3) regression (or classification). Feature extraction plays a significant role for depression recognition in videos. Mining a discriminative feature descriptor is crucial and meaningful for depression recognition and estimation. From the perspective of feature extraction, the technologies can be roughly divided into hand-crafted features, and deep-learned features.

Hand-crafted features utilize domain knowledge to design features that are closely related to the symptoms of depression [2], [3]. Though hand-crafted feature representations have been considered to obtain superior performance for assessing the severity of depression, there exist the following issues reported by researchers. Firstly, exploiting hand-crafted features is time-consuming as they need task-specific knowledge. Local Binary Patterns from three orthogonal planes (LBP-TOP) [4] is typically and computationally heavy. Secondly, hand-crafted features are criticized as lacking significant information closely related

Table 1: BDI–II Score Ranges and Depression Severity.

BDI–II Score Ranges	Depression Severity
None or minimal	0 – 13
Mild	14 – 19
Moderate	20 – 28
Severe	29 – 63

to depression patterns [5].

Recently deep-learned features using convolutional neural networks (CNN) has been widely used for deep feature representation and has performed well in depression recognition and analysis [6]. DepressNet [6] is a novel framework, to learn a depression representation with explanation. A selected CNN model (e.g., AlexNet [7], GoogleNet [8], VGG-Net [9], etc.) is pre-trained on the large-scale facial image dataset [10], followed by fine-tuning on the AVEC2013 [11] and AVEC2014 [12] datasets. The performance surpasses most of the video-based depression recognition methods. In [13], a combination of Recurrent Neural Network (RNN) [14] and 3D convolutional neural network (C3D) [15] is used to learn sequential representation of the spatio-temporal features at two different scales from the face regions. Comprehensive experiments on the two depression databases (i.e., AVEC2013 and AVEC2014) demonstrate that the C3D method is promising. In [6] and [13], the authors adopt the pre-trained deep model to fine-tune on the two depression databases for depression estimation. And the scheme cannot be considered an end-to-end method for depression recognition.

In general, depression recognition is a regression or classification issue from the perspective of machine learning. AVEC2013 and AVEC2014 aim at predicting the depression scores, the Beck Depression Inventory–II (BDI–II [16], Table 1) recorded video of a subject. It is suggested that most non-verbal behaviours in human interaction are around the facial region [17]. For visual-based depression estimation, salient region of the face can be used to predict the severity of depression, in particular the patch of eye region is crucial, as suggested in [6].

In this paper, the focus is on exploring the technologies based on the facial appearance for depression recognition. To overcome the aforementioned problems, we propose a novel framework, termed Deep Local-Global Attention Convolutional Neural Network (DLGA-CNN), for depression recognition and analysis using facial images/videos. More specifically, as shown in Figure. 1, the DLGA-CNN consists of three components: 1) deep-learned feature extraction module Depressed-CNN, 2) Local-Global-Attention-based Convolutional Neural Network (LGA-CNN), and 3) Weighted Spatial Pyramid Pooling (WSPP) module. The Depressed-CNN module extracts feature map representation from the entire image which can help to learn deep characteristic patterns of depression. The LGA-CNN represents the global and local attentive features of the feature map. Global features aim to describe the ensemble of special patterns of depression while local features concentrate upon capturing specific patterns on patch regions, which can mine discriminative characteristic on salient patches of the feature maps. The WSPP module 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.

The key contributions of the present paper can be summarized as follows:

1. We propose an end-to-end framework DLGA-CNN that effectively captures the facial dynamics as a non-verbal measure for estimating the severity of depression scale.
2. To encode the robust feature representations, a deep-learned convolutional feature extraction network named Depressed-CNN is designed. Valuable and discriminative features helpful for depression analysis are retained with the Depressed-CNN.
3. A CNN with self-attention network, termed LGA-CNN, efficiently describes the discriminative patterns from the faces. By adopting the atten-

tion mechanism, LGA-CNN can automatically retain the valuable characteristic and filter the redundant information of the face.

4. Extensive experiments on the two databases (i.e., AVEC2013 and AVEC2014) have been conducted to compare with other visual-based depression recognition methods. The results demonstrated the effectiveness of the proposed method.

The rest of the present paper is organized as follows. Section 2 briefly discusses previous work on visual-based depression analysis and recognition. Section 3 details the proposed methods. Section 4 introduces the databases adopted and the experimental results. Conclusions and future works are discussed in Section 5.

2. Related Work

A great number of approaches for automatic depression analysis have been designed based on AVEC2013 and AVEC2014. For audiovisual-based depression recognition, the step of feature extraction plays a vital part for estimating the depression scale. In recent years, hand-crafted visual features have been widely used and proved efficient for depression recognition. In the following, we introduce the visual-based hand-crafted approaches with the description of deep-learned features to predict the scale of depression.

In [2], Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) features are captured with sparse coding, and then represented by discriminate mapping and decision level fusion method to generate high level features. Support vector regression (SVR) is adopted to assess the scale of depression severity.

In [18], a novel dynamic temporal feature, termed Median Robust Local Binary Patterns from Three Orthogonal Planes (MRLBP-TOP), is used for presenting the temporal facial dynamics expressions. To make a feature vector representation, the Dirichlet Process Fisher Vector (DPFV) learns richer and more compact intermediate representation of MRLBP-TOP features on

sub-sequences. Then for each video sample, a discriminative representation is created with statistical aggregation approaches.

The above-mentioned frameworks mainly focused on hand-crafted features. They outperform the majority of the-state-of-the-art methods with lower errors rates and they performed the experiments with the two depression databases (i.e., AVEC2013 and AVEC2014). In the following section, we provide a brief review over the deep-learned methods related to video-based depression recognition.

Zhu et al. [19] train a 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 [20], an AI system is designed to estimate the scale of depression. The system can fuse the complementary pattern between the hand-crafted and deep-learned features. For visual cues, deep-learned features are extracted, which include some discriminative information related to depression. For audio cues, they extract Spectral Low-level Descriptors (LLDs) and Mel-frequency cepstral coefficients (MFCCs) to capture vocal expression from audio clips. Temporal movements on the different feature space are described by Feature Dynamic History Histogram (FDHH).

In [6], the authors propose DepressNet, a deep regression network for predicting the severity of depression from single images. Depression activation map (DAM) is adopted to represent salient regions of the facial image to determine the scale of depression. Meanwhile, they also design a multi-region DepressNet to model the different patterns of different regions to improve overall results. Extensive experiments conducted on AVEC2013 and AVEC2014 databases, and the performance demonstrated that the proposed approach surpasses the most of the state-of-the-art visual-based depression recognition methods.

In [21], they extract local-global features of convolutional 3D networks to boost the ensemble performance. The proposed network is equipped with 3D global average pooling to represent spatiotemporal patterns for depression detection. Experimental results show that fusing the local and global features of

C3D networks obtains promising performance.

Song et al. [22] propose to automatically extract human behaviour primitives as the low-dimensional descriptor of each frame. Two spectral feature representations, i.e., spectral heatmaps and spectral vectors, are proposed to capture multi-scale patterns related to depression. The authors fed the two spectral representations to CNN and Artificial Neural Networks (ANNs) for depression prediction. This is the first work on representing human behaviour primitives for depression recognition.

In [23], a two-stream spatiotemporal framework for depression recognition is proposed. Also, the authors propose a temporal median pooling (TMP) method to generate temporal fragment features. Experiments on AVEC2013 and AVEC2014 database demonstrated that the proposed framework obtained comparable performance for depression recognition.

For deep-learned features, the existing approaches are first trained on a large database (e.g., CASIA WebFace Database, etc.), and fine-tuned on AVEC2013 and AVEC2014 databases [19, 20, 6, 21]. Meanwhile, in [22] and [23], the authors train the deep models from scratch for depression recognition. These approaches have one thing in common that they cannot be considered as an end-to-end scheme for depression recognition. More importantly, there is no guideline as to which network structure is best-suited for modelling the representations of depression (i.e., the depth of the network, etc)?

In our study, we intend to exploit a discriminative and an informative representations for depression prediction, which can directly predict the depression scale from facial images. More specifically, our approach targets an end-to-end framework from facial images to direct prediction of the severity of depression.

Finally, we highlight some difference of our proposed framework with existing methods. Firstly, to the best of our knowledge, our proposed framework is the first study to adopt attention mechanisms to model discriminative patterns closely related to depression recognition. Secondly, in terms of aggregation local features to global, existing methods consider the local and global features as dependent steps. We propose to integrate the different feature representations

steps into an end-to-end scheme over the video clip for depression recognition and prediction.

3. The Proposed Approach

In this Section, we first provide an outline of the proposed framework in sub-section 3.1, and then introduce the feature extraction in sub-section 3.2, followed by detailed LA-CNN and GA-CNN in sub-section 3.3 and 3.4. In sub-section 3.5, we briefly describe WSPP to transform and extract final feature representations.

3.1. Framework Overview

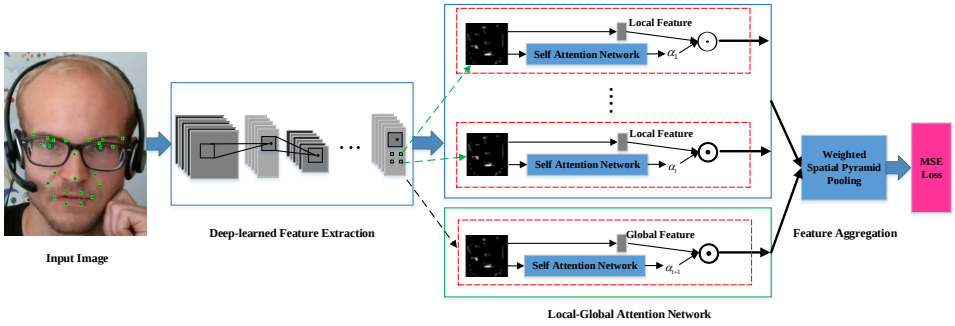


Figure 1: The pipeline of the proposed architecture for the diagnosis of depression.

The end-to-end deep depression recognition framework is illustrated in Figure. 1. The facial region is cropped from the video clips with OpenFace toolkit [24]. After that, a typical CNN is implemented for feature extraction, which obtain the discriminative facial feature maps. To filter redundant features, different self attention network over the local and global feature maps are introduced. To make scale-invariant feature representations over multi-scale feature maps, WSPP is adopted. Lastly, two fully connect layers and a mean square error (MSE) loss layer are utilized to predict the depression severity. In the following, we detail each component of the framework.

Table 2: The network configuration of Depressed-CNN. Conv denotes the convolutional operation of CNN. CMP is the Channel Max Pooling operation.

Name	Input	Operation	Kernel	Output
Conv1	$224 \times 224 \times 3$	Convolution	3×3, ReLU	$224 \times 224 \times 64$
Pool1	$224 \times 224 \times 64$	Pooling	2×2	$112 \times 112 \times 64$
Conv2	$112 \times 112 \times 64$	Convolution	3×3, ReLU	$112 \times 112 \times 128$
Pool2	$112 \times 112 \times 128$	Pooling	2×2	$56 \times 56 \times 128$
Conv3	$56 \times 56 \times 128$	Convolution	3×3, ReLU	$56 \times 56 \times 256$
Pool3	$56 \times 56 \times 256$	Pooling	2×2	$28 \times 28 \times 256$
Conv4_1	$28 \times 28 \times 256$	Convolution	3×3, ReLU	$28 \times 28 \times 512$
Conv4_2	$28 \times 28 \times 512$	Convolution	3×3, ReLU	$28 \times 28 \times 512$
CMP	$28 \times 28 \times 512$	Pooling	4×4	$28 \times 28 \times 128$
Con	CMP+Conv4_2	Concatenate	/	$28 \times 28 \times 640$

3.2. Deep Learned Feature Extraction

CNN has been widely adopted and proved efficient for feature extraction in affective computing field, such as facial expression recognition [25], depression recognition [20], etc. To deal with small size dataset, the proposed framework for depression analysis is inspired by the work described in [26] with shallow architectures. Because of the shallow architectures can model a discriminative representation for predicting the severity scale of depression. For deep-learned feature extraction, the proposed configuration of Depressed-CNN architecture is illustrated in Table 2. Meanwhile, Depressed-CNN is also detailed in Fig. 2 with blue rectangle.

The deep-learned features from the Depressed-CNN are extracted as follows. The size of input facial image is 224×224 with 3 color channels (RGB). Inspired by [9], for all the convolutional layers, we also adopt filter with a small 3×3 kernel to encode the notion of left/right, up/down for extracting the spatiotemporal information. After three ordinary convolutional and max pooling operation, the size of Conv4_1 is $[512 \times 28 \times 28]$. In our work, max pooling is performed

over a 2×2 window with stride=2. To avoid information loss and retain the discriminative feature, we use channel max pooling (CMP) to pool the feature map in the channel direction with kernel size 4, stride=4 and pad=0. The output of a three-dimension feature map $[128 \times 28 \times 28]$ is obtained. Such construction is to have the common max-pooling calculate the maximum value in the spatial direction, while the channel max pooling to calculate the maximum value of the channel (RGB) direction. Moreover, to make a robust feature representation, the Conv4_2 feature map is generated over Conv4_1. After that, the Conv4_2 feature map and the CMP are concatenated to obtain the final feature with the size of $[640 \times 28 \times 28]$.

3.3. Local Attention-based CNN (LA-CNN)

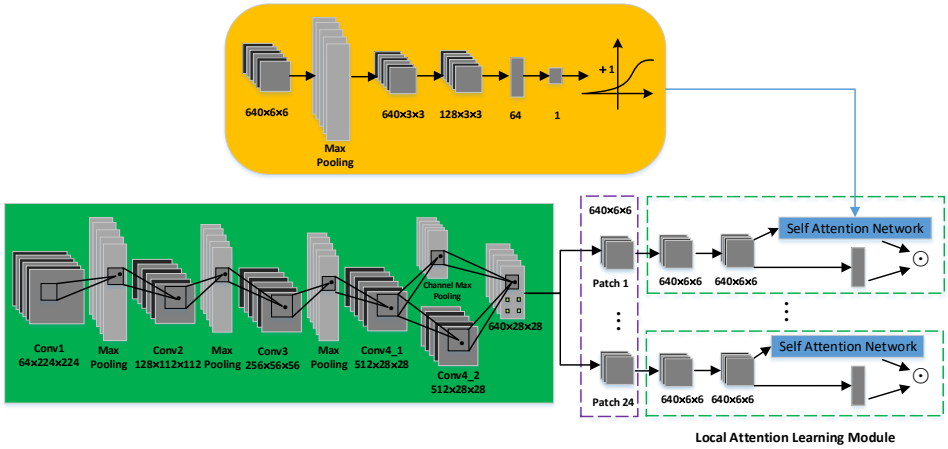
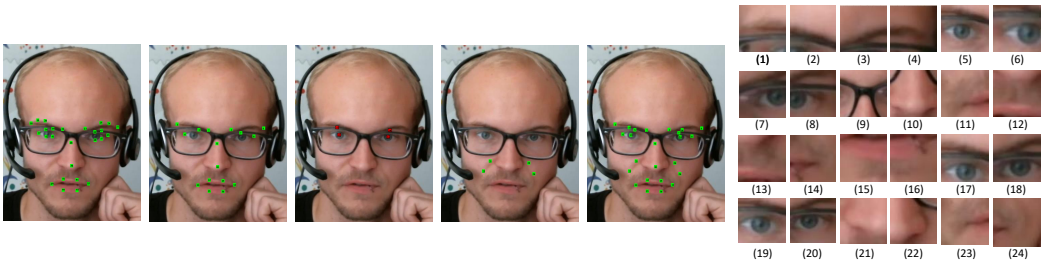


Figure 2: The detailed illustration of the LA-CNN.

Different parts of the facial image have respective contribution for depression recognition. Inspired by [27], we crop the facial region into different patches to capture discriminative feature representation for depression analysis. LA-CNN includes two steps: patch generation and salient features capture. In the following, we describe the two steps in detail.



(a) (b) (c) (d) (e) (f)

Figure 3: Patch generation of the face. (a) 68 facial landmarks are detected using the OpenFace toolkit. 26 points are selected from the 68 points, which covered the main facial appearance. The index of the selected points is 18, 19, 20, 37, 38, 39, 41, 42, 22, 23, 25, 26, 27, 44, 45, 46, 48, 47, 28, 30, 49, 51, 59, 53, 55, 57. (b) To further compute the number of facial points, 16 points are re-selected and re-computed from the 26 points. (c) We take four pairs of facial landmark points (20, 38), (42, 41), (25, 45), (47, 48) and compute the midpoint of them. (d) We take out two pair of points (18, 59), (27, 57) and compute the midpoint of them. The index of midpoint are 21, 22. After the computation of the distance between the mouth corner and the target point, the index of them is 23, 24. (e) 24 facial landmarks are generated. (f) 24 patches are cropped based on the facial landmarks from the facial region.

3.3.1. Patch Generation

For depression analysis from facial region, local patches may have discriminative characteristic for estimating the severity of depression. As a common deep-learned feature extraction scheme, a limitation of CNN is to learn geometric transformations. As some specific facial muscles (areas) are informative for expression recognition closely related to depression [28], we crop the facial region into discriminative patches (e.g., around mouth and eyes). Also, in [29] and [30], the authors consider that multi-view information of salient regions is important for image retrieval. To this end, we propose to adopt different patches to capture discriminative representations for depression recognition. In this work, the OpenFace toolkit is adopted to detect the facial region and to align the each frame of the video sequences. After that, 68 facial landmark points are detected (Fig. 3). To locate the efficient facial patches that are

closely related to depression, we select 16 points from 68 points that cover the discriminative patches (i.e., eyes, mouth and nose) of the facial region. Then eight points are re-computed that covered over eyes and facial cheeks. A total of 24 facial patches are extracted and the detailed procedure is shown in Fig. 3.

Steps of the processing chain:

(a) The OpenFace toolkit is utilized for facial region detection and face alignment in each image of the video sequence. We select 26 points that cover the main facial regions, i.e., mouth, eyes and nose. The selected points are indexed: 18, 19, 20, 37, 38, 39, 41, 42, 22, 23, 25, 26, 27, 44, 45, 46, 48, 47, 28, 30, 49, 51, 59, 53, 55, 57.

(b) We take four pairs of the facial landmark points (20, 38), (42, 41), (25, 45), (47, 48). Finally, 16 points are re-computed from the 26 points.

(c) The midpoint of each pair points is computed.

(d) To explore the region around the mouth area, we use two pairs of points (18, 59) and (27, 57) and compute the midpoints. The indexes of the midpoint are 21, 22. Then we compute the two points that have the same distance from the mouth corners. For the coordinate of the left target point, it is defined as $(U, V) = (U_{left} - 16, V_{left} - 16)$. While for the coordinate of the right target point, it is $(U, V) = (U_{right} - 16, V_{right} - 16)$. The distance between the mouth corner and the target point is computed and the indexes are 23 and 24.

(e) Then, we select 24 facial landmark points that cover the main facial regions.

(f) According to position of the 24 points, 24 patches are cropped. As shown in Fig. 3f, the patches are generated from the facial image. However, in our work, patch generation is performed on feature maps to obtain the salient regions.

3.3.2. Salient Features Capture

As shown in the green box of Fig. 2, the patch generation process is carried out using the feature maps of CNN rather than the facial images. This is to maximize the use of convolutional operations and amplify the receptive fields of

neurons. This has reduced the size of model. The output of feature extraction procedure is the feature maps $[640 \times 28 \times 28]$. For patch generation, we obtain 24 local regions with size of $[640 \times 6 \times 6]$.

To obtain the salient features, we use the Local Attention Learning (LAL) module with the LA-CNN to automatically learn the characteristic of facial patches. The LAL module is shown in the yellow box in the middle of Fig. 2 with the two green dashed rectangles. In each patch-specific LAL module, after patch generation, the generated feature maps are input into two convolutional layers. After that, the second feature maps are input into two branches. The first branch treats the feature maps as the vector-level local feature. While for the second branch, a self attention net is adopted to concentrate on the discriminative representation regions from the spatial patches. Then the patch feature are re-computed with the weight vector.

Formally, let $\mathbf{Pa}_j$ be the j_{th} patch of the feature map with the size of $[640 \times 6 \times 6]$. $\hat{\mathbf{Pa}}_j^1 = f(\mathbf{Pa}_j)$ is the first feature map in the top green dashed rectangle in Fig. 5 with the size of $[640 \times 6 \times 6]$. After a 1×1 filter, the second feature map is $\hat{\mathbf{Pa}}_j^2 = f(\mathbf{Pa}_j)$ with the size of $[640 \times 6 \times 6]$. f denotes the convolution operation of CNN architecture. Then we feed the second feature map into the two branches.

The first branch transforms the second feature map into a vector of local feature. Let φ_j be the local feature which take the second feature map as input. The φ_j can be written as:

$$\varphi_j = \varphi(\hat{\mathbf{Pa}}_j^2) \tag{1}$$

where φ is the vector transform operation.

The second branch is the self attention net, which includes one max-pooling operation, one convolution operation, two fully connect layers, and a sigmoid operation. The sigmoid function is used for restricting the range of the output α_j from 0 to 1, where 0 denotes the irrelevant patch while 1 represents the closely important patch of depression. The learned weight α_j can be defined as:

$$\alpha_j = \omega_j(\hat{\mathbf{Pa}}_j^2) \tag{2}$$

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

Lastly, after the operation of two branches, the weight α_j is performed on the local feature φ_j to generate the discriminative feature:

$$\rho_j = \alpha_j \cdot \varphi_j \tag{3}$$

The output feature will contain discriminative representation related to depression. More specifically, each LAL module is weighted by the weights that automatically learned by the self attention net. For the LA-CNN, it is an end-to-end local depression framework with the following components: deep-learned feature extraction by CNN, patch generation, and attention mechanism. The patch-based characteristic learned by our deep model is able to discover the visual depression pattern on facial region, with which automatic depression estimation is possible.

3.4. Local-Global Attention-based CNN (LGA-CNN)

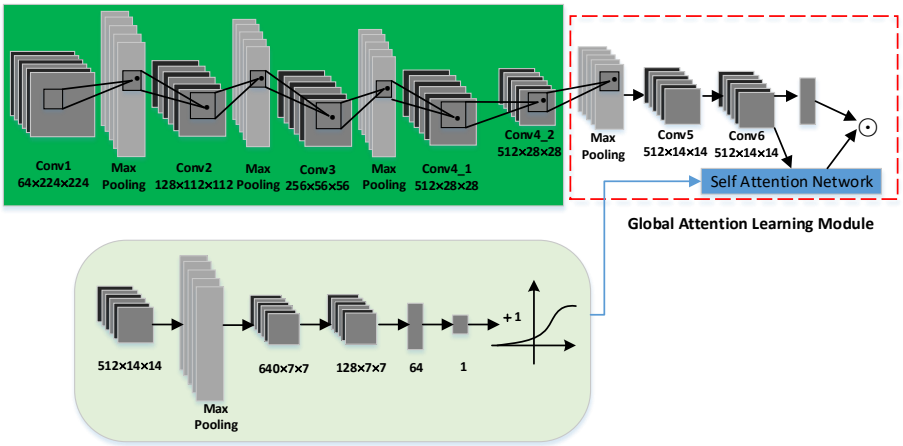


Figure 4: The detailed illustration of the GA-CNN.

This sub-section is to describe the GA-CNN, followed by the introduction of the LGA-CNN architecture.

As described above, LA-CNN can automatically learn discriminative characteristic with attention technology for depression analysis. However, the patches

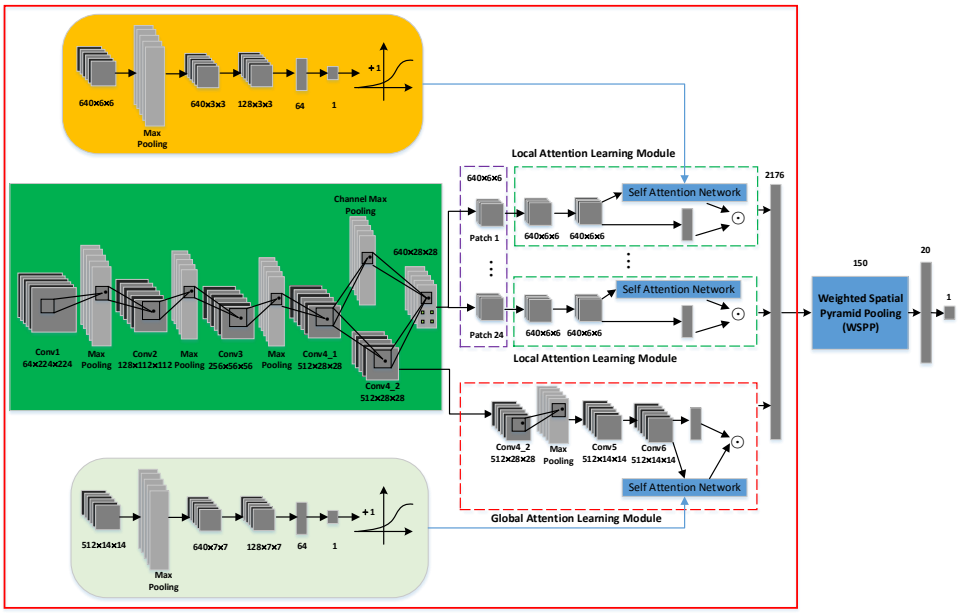


Figure 5: The detailed illustration of the LGA-CNN-WSPP.

of LA-CNN may miss some complementary information included of facial images and global semantic information about the pattern of depression the full facial image may have. To improve the overall performance and learn the deep semantic information, a LGA-CNN is represented.

For the part of feature extraction, we use the same operation as LA-CNN. In our implementation, we propose to use Global Attention Learning (GAL) module to represent the global semantic information for depression recognition (the red dashed rectangle in Fig. 4). The GAL module consists of one max-pooling operation, one convolution operation, and two branches operation. Before feeding into the two branches, the feature map Conv6 can be described as g with the size of $[512 \times 14 \times 14]$. The first branch transforms the feature maps g into a vector of global feature. Let ψ be the global feature, and written as:

$$\psi_{j+1} = \varphi(g) \tag{4}$$

where φ is the vector transform operation.

The second branch is the self attention net, which consists of one max-pooling operation, one convolution operation, two fully connect layers, and a sigmoid operation. The learned weight α_{j+1} can be defined as:

$$\alpha_{j+1} = \omega_{j+1}(g) \tag{5}$$

where α_{j+1} is a scalar, ω_{j+1} represent the operations of self attention net.

Finally, the weight α_{j+1} is weighted on the global feature φ_{j+1} to get the informative feature ρ_{j+1}:

$$\rho_{j+1} = \alpha_{j+1} \cdot \varphi_{j+1} \tag{6}$$

Moreover, to encode the complementary representations within the proposed LA-CNN and GA-CNN, we propose an end-to-end architecture, termed LGA-CNN (the red rectangle in Fig. 5), to concatenate them into the ensemble feature to estimate the depression scale.

3.5. WSPP

The main goal of this work is to assess the severity of depression. To enhance the indicators for severity of depression, the deep representation has to capture discriminative facial features in different scales, hence a significant number of instances in all its natural cases are needed in training. The spatial sources of natural variations of face includes the viewpoint (angles), the location, and the size of the facial region.

In our work, we use the local and global deep learned features with attention mechanism to directly capture the discriminative information for depression recognition. However, the head movement may result in the size variations of the face in an image sequence. Therefore, the variations may lead to the face-blurring which will affect the resolution of the facial image.

To obtain the scale-invariant representation, a WSPP layer is adopted to represent multi-scales on top of the output of LGA-CNN. The idea of WSPP is to segment the feature map into different divisions from finer to coarser scales, finalized with an aggregation of local features. The WSPP layer can improve scale-invariance and reduce the problem of overfitting. In the present paper, we

utilize the original weighting definition on the spatial pyramid kernel in [31]. The features at finer resolutions are associated with a heavier weight, and coarser resolutions with a lower weight. Fig. 6 details the feature representation step of WSPP. In our task, the shape of the output of LGA-CNN is $2048 \times 1 \times 1$.

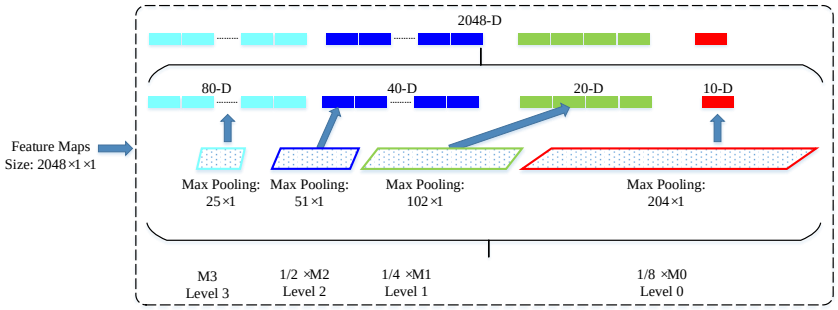


Figure 6: The detailed illustration of the WSPP.

In every spatial bin, we use max-pooling to pool the responses of each filter. The output of WSPP is the sum of the total number of bins. In our architecture, the output of LGA-CNN has a shape of [batch_size, 2048], we obtain a final feature vector with the size 150D. The window sizes of WSPP is 25×1, 51×1, 102×1, 204×1 and their corresponding stride is 25×1, 51×1, 102×1, 204×1, respectively. The fixed-dimensional vectors are the feed into the fully-connected layer.

As a deep network, the loss function play a significant role in the final regression. In our task, depression analysis can be regarded as a regression issue. Therefore, we use Euclidean loss as the loss function, which is thought suitable for our work. Formally, the Euclidean loss function L calculates the sum of squared differences between predicted and ground truth values, which can be expressed as:

$$L = \frac{1}{2M} \sum_{j=1}^{M} \|\hat{p}_i - p_i\|^2 \tag{7}$$

where M denotes the number of samples, $\hat{p_j}$ is the output from the architecture, and p_j is the label. The detailed illustration of the LGA-CNN-WSPP in Figure.

5.

4. Experiments

This section gives an introduction to the experimental evaluations of the proposed scheme for depression recognition. Comprehensive experiments are performed on the two depression databases (i.e., AVEC2013 and AVEC2014) to evaluate the proposed framework. The goal of the experimental works are as follows:

1. Investigate the performance of the two branches in LGA-CNN, i.e., the LA-CNN and the GA-CNN.
2. Evaluate the performance of proposed framework for depression recognition.
3. Estimate the expandability of the proposed framework on the two databases.

In sub-section 4.1, the AVEC2013 and AVEC2014 depression databases are introduced. The details of experiments are presented in sub-section 4.2. In sub-section 4.3, we show and discuss the performance of the proposed method for assessing the scale of depression.

4.1. AVEC2013 and AVEC2014 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 samples.

In AVEC2013 depression corpus, there are totally 150 video clips from 82 subjects. The audioviusal 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

In this work, OpenFace toolkit is adopted to detect the facial region and align the landmarks localization on on AVEC2013 and AVEC2014 depression databases. The generated facial images have a size of 224×224 with RGB color channels. When we perform the experiment on the two depression databases, it is not guaranteed that the facial region is detected by the OpenFace toolkit in all frames. Therefore, it has to be checked manually frame by frame on the two databases to make sure the detected facial region for depression prediction.

For LA-CNN-WSPP of DLGA-CNN, the feature maps extracted from the Depressed-CNN are split into local regions and equipped with 24 LAL modules as well as transformed with WSPP for non-variant feature aggregation to predict the depression scale. For GA-CNN-WSPP of DLGA-CNN, the feature maps extracted from the Depressed-CNN are represented as a whole architecture and equipped with a GAL module, and used WSPP for non-variant feature aggregation to predict the severity of depression. In our current implementation, we directly train the whole framework into an end-to-end manner rather than use the pre-trained architecture for predicting the severity of depression.

The Networks are trained by TensorFlow deep learning toolbox [32]. To obtain fast convergence, the adaptive moment estimation (ADAM) [33] optimizer is used for training of our proposed method, which has been considered as an useful approach to model and evaluate deep architecture on small databases. To construct an end-to-end framework for depression recognition, the parameters of the network are set empirically and using reference from another studies [26]. Therefore, we train the network from scratch with different combination of parameters and optimize it with many attempts. For the ADAM parameters, the default value of β_1, β_2, and ε is 0.9, 0.999, and 1e-8, respectively. To improve the ability of generalization, L2 weight decay (factor 0.0005) is adopted in the

all experimental steps. The learning rate is set to 0.000001. The batch size is 32. We conduct the experiments with two Titan-X GPU (each with 12G memory). The number of iterations is empirically set to 50k and it will take 5 hours to train the deep depression recognition model.

4.2.2. Evaluation Metrics

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. 8 and Equ. 9, respectively, where N denotes the number of subjects, p_i represents the BDI-II score, and $\tilde{p}_i$ is the output value of the j-th subjects.

$$MAE = \frac{1}{N}\sum_{j=1}^{N}|p_j - \tilde{p}_j| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(p_j - \tilde{p}_j)^2} \tag{9}$$

4.3. Experimental Results

In this section, we first consider the overfitting and undertraining issues in the training process. Second, an ablation study is first performed to evaluate the performance of each part of the proposed framework. Third, the proposed framework is further compared with several video-based depression recognition methods to show its excellent performance.

4.3.1. Training of Deep Models

In our study, an early-stop strategy is adopted to avoid overfitting and undertraining. Specifically, if the loss of model training is no longer decreasing in two hundred iterations, the training is stopped. If the loss continues to decrease during the training step with an decreasing accuracy with the testing set, we consider that overfitting happened and stop the training. On the other hand, to avoid undertraining, we attempt to train the model from scratch many times

Table 3: Performance of different combinations of the proposed framework on the test set of AVEC2013.

Model Setting	Evaluation Metrics	
	RMSE	MAE
LA-CNN (CMP-)	8.84	7.26
LA-CNN (CMP+)	8.80	7.22
A1: LA-CNN (CMP-) + WSPP	8.60	6.90
B1: LA-CNN (CMP+) + WSPP	8.44	6.63
GA-CNN	9.14	7.51
C1: GA-CNN + WSPP	9.07	7.44
D1: LA-CNN (CMP-) + GA-CNN	8.80	7.23
E1: LA-CNN (CMP+) + GA-CNN	8.74	7.10
F1: LA-CNN (CMP-) + GA-CNN + WSPP	8.61	6.90
G1: LA-CNN (CMP+) + GA-CNN + WSPP	8.39	6.59

and evaluate the performance of the trained models with other existing deep learning methods for depression recognition and analysis.

4.3.2. Ablation Study

The effectiveness of each component of the proposed framework is evaluated on the AVEC2013 and AVEC2014 databases. The model setup by "LA-CNN (CMP+)", "LA-CNN (CMP-)" represents that the CMP technology is used or not on LA-CNN, respectively. The model setting "GA-CNN" denotes that only global attention method is adopted for depression recognition. The rest of model setting in Table 3 and 4 combines two or three individual components to capture the complementary information between them. For instance, G1 and G2 represent the model that combine three individual feature extraction components.

Table 3 illustrates the recognition results on the test set of AVEC2013 by RMSEs, along with the combination of individual components. As can be seen, the combination G1 obtains the best performance, the RMSE of 8.39, followed

Table 4: Performance of different combinations of the proposed framework on the test set of AVEC2014.

Model Setting	Evaluation Metrics	
	RMSE	MAE
LA-CNN (CMP-)	8.81	7.25
LA-CNN (CMP+)	8.79	7.22
A2: LA-CNN (CMP-) + WSPP	8.60	6.90
B2: LA-CNN (CMP+) + WSPP	8.42	6.62
GA-CNN	9.11	7.50
C2: GA-CNN + WSPP	9.00	7.41
D2: LA-CNN (CMP-) + GA-CNN	8.70	7.02
E2: LA-CNN (CMP+) + GA-CNN	8.66	6.93
F2: LA-CNN (CMP-) + GA-CNN + WSPP	8.56	6.81
G2: LA-CNN (CMP+) + GA-CNN + WSPP	8.30	6.51

by the combination B1, the RMSE is 8.44.

While for AVEC2014 database, we also conduct various experiments to validate the effectiveness of the proposed framework. From Table 4, one can see that we obtain the same observations as AVEC2013. As illustrated in Table 4, the results after combining the different components obtained the RMSE 8.30 on AVEC2014 database. Furthermore, from the two tables, one can notice that the different models based on LA-CNN perform better than GA-CNN and C1 models. For instance, the RMSE of GA-CNN is 9.11, while the lowest RMSE is 8.42 of B2. Moreover, both results perform better than those adopting each individual components. The performances on the two depression databases show that the capability of the proposed method to assess the severity scale of depression from video sequences. This observation demonstrates that by combining the individual component, the overall performance can be further promoted than using an individual component, which further means the necessity of integrating the local and global models for depression recognition and analysis.

Table 5: Performance of different learning rate of the proposed framework (G1) on the test set of AVEC2013.

Learning Rate	RMSE	MAE
0.000001	8.39	6.59
0.00001	8.59	6.90
0.0001	8.78	7.21
0.001	9.10	7.50
0.01	9.28	7.60
0.1	9.36	7.65

Table 6: Performance of different learning rate of the proposed framework (G2) on the test set of AVEC2014.

Learning Rate	RMSE	MAE
0.000001	8.30	6.51
0.00001	8.54	6.89
0.0001	8.80	7.22
0.001	9.08	7.49
0.01	9.26	7.48
0.1	9.35	7.65

We adopt various parameter settings to train and evaluate the proposed framework for depression recognition. To make fair comparisons, we only adopt G1 and G2 to evaluate the models for depression recognition. We change the learning rate from 0.000001 to 0.1, with other parameters fixed. As shown in Tables 5 and 6, increasing learning rate reduces performance. As larger learning rates result in inferior performance, minimal learning rate is suitable for the proposed framework.

As shown in Tables 5 and 6, we can see that increasing learning rate from 0.000001 to 0.1 reduces the performance, while larger learning rate results in inferior performance, which demonstrate the minimal learning rate is suitable

for the proposed framework to learn the important patterns closely related to depression.

4.3.3. Analyzing of Computational Complexity

The speed of the model is evaluated on an Intel® Core™ i7-6700 CPU @ 3.40 GHz with 32 GB memory using TensorFlow platform. As mentioned above, an end-to-end deep recognition model is trained in 5 hours. When we test the performance of the model, a single image took 2.12 seconds. The computational complexity of a video clip depends on its length. Though concurrent works for depression recognition do not always consider the computational complexity, with an exception in [18]. The low computational complexity of our proposed method further illustrates its effectiveness.

4.3.4. Comparison With Previous Works

To further demonstrate the ability of the proposed framework, we finally compare our depression recognition methods, using the proposed method combining different approaches, with those of previous methods using visual-based features.

The performance of the two public depression databases are demonstrated in Table 7 and 8, respectively. On the AVEC2013 database, as presented in Table 7, our proposed framework yields comparable performances than most of the video-based results, except for the approach of [6], [21], [36]. The baseline work adopts hand-crafted feature representations generated with the LPQ texture descriptor [11]. It is noted that the proposed method outperforms the baseline system by margin (e.g., 5.35 in term of RMSE). The study [6] proposes a novel deep regression framework namely DepressNet to model a robust representation by visual explanation of the depression severity. The authors first pre-train the various deep models (e.g., GoogleNet, VGG-Net, etc.) combined with a global average pooling layer on the large scale facial image database, and fine-tune the models on AVEC2013 and AVEC2014 databases. Furthermore, they also adopt the different facial region to fuse the complementary information to promote

Table 7: Performance of different architecture for visual-based automatic depression diagnosis on the test set of AVEC2013.

Methods	**RMSE**	**MAE**
Baseline [11]/ LPQ, SVR	13.61	10.88
Cummins et al. [34]/ STIP and PHOG, SVR	10.45	N/A
Meng et al. [35]/ EOH and LBP, PLSR	11.19	9.14
Wen et al. [2]/ LPQ-TOP, SVR	10.27	8.22
Zhu et al. [19]/ Optical Flow, 2D-CNN	9.82	7.58
Mohamad et al.[13]/ C3D, RNN	9.28	7.37
He et al. [18]/ MRLBP-TOP, DPFV, SVR	9.20	7.55
Md et al. [23]/ LSTM	8.93	7.04
Zhou et al. [6]/ 2D-CNN	8.19	6.30
Melo et al. [21]/ C3D	8.26	6.40
Melo et al. [36]/ ResNet-50	8.25	6.30
Proposed Approach	8.39	6.59

the overall depression recognition results. In comparison with the results shown in [6], our method has not improved the accuracy of depression recognition. There is a potential reason that two stage framework (i.e., pre-train, fine-tune) can effectively use the advantage of them for depression recognition. More importantly, the large webface database can provide a large number of samples to pre-train the deep models to learn the various patterns related to depression. In [21], the authors propose to adopt convolutional 3D networks that are pre-trained on CASIA dataset to represent spatio-temporal feature representations, where a deep architecture is adopted for estimating the severity of depression. Meanwhile, they also adopt 3D Global Average Pooling (3D-GAP) to model the local and global features for predict the depression levels. In [23], the authors propose a two stream framework for depression recognition. In the framework, a volume local directional number (VLDN) feature and the spatial features by the Inception-ResNet-v2 network are combined to learn the spatialtemporal

Table 8: Performance of different architecture for visual-based automatic depression diagnosis on the test set of AVEC2014.

Methods	**RMSE**	**MAE**
Baseline [12]/ LGBP-TOP, SVR	10.86	8.86
Sidorov et al. [37]/ LGBP-TOP, SVR	10.83	8.32
Jan et al. [38]/ EOH, LBP and LPQ, PLSR	10.50	8.44
Kaya et al. [39]/ CCA, ELM	10.27	8.20
Zhu et al. [19]/ Optical Flow, 2D-CNN	9.55	7.47
Mohamad et al.[13]/ C3D, RNN	9.20	7.22
He et al. [18]/ MRLBP-TOP, DPFV, SVR	9.01	7.21
Dhall et al. [4]/ LBP-TOP, SVR	8.91	7.08
Md et al. [23]/ LSTM	8.78	6.86
Zhou et al. [6]/ 2D-CNN	8.39	6.21
Melo et al. [21]/ C3D	8.31	6.59
Melo et al. [36]/ ResNet-50	8.23	6.13
Proposed Approach	8.30	6.51

representation patterns related to depression. By comparing the works [6], [21], [36], the RMSEs are not surpass of them, our method is trained from scratch that is an end-to-end scheme for depression recognition.

For AVEC2014, as illustrated in Table 8, our method obtains comparable results than the most of the video-based depression recognition methods, with 6.59 MAE and 8.39 RMSE on the test set. The baseline study adopts hand-crafted features generated with the LGBP-TOP descriptor [12], the RMSE is reduced by 23%. Moreover, these results of different components are better than those obtained on AVEC2013 database.

In Tables 7 and 8, the method of our previous work [18] propose a spatial-temporal dynamic feature descriptor MRLBP-TOP and DPFV for feature extraction and aggregation. Results demonstrate that the proposed hand-crafted features as well as feature aggregation method obtain the higher RMSEs, when

compared with the method proposed in the present paper. From these results, a conclusion can be reached that the proposed method on AVEC2013 and AVEC2014 databases can automatically learn local and global characteristic information of facial region, and outperform the most of the state-of-the-art methods for depression recognition. This further demonstrates the effectiveness of the proposed method for depression recognition and analysis.

5. Conclusion and Future Works

In the present paper we adopt CNN with attention mechanism to design an integrated end-to-end framework for video-based depression recognition. We argue that such an efficient capability is significant for capturing the characteristic pattern of depression "encoded" in facial regions. Specifically, a novel framework termed DLGA-CNN is proposed. The framework includes two branches: LA-CNN and GA-CNN. LA-CNN only focus on the local patches. GA-CNN learns the global patterns from the entire facial region. To capture the complementary information between the two branches, LGA-CNN is proposed. Lastly, to further obtain the informative depression patterns, WSPP is also used to learn the representations of the final features. Comprehensive experiments on the two public depression databases (i.e., AVEC2013 and AVEC2014) have shown that the capability of the proposed framework, when compared with the most of the state-of-the-art video-based depression recognition approaches.

In the future, we will consider different features using deep learning for depression recognition. Additionally, we will explore more explicable representation patterns and more robust regression models with discriminative deep network. More importantly, the proposed deep learning based technology can help the clinicians to assess depressed subjects. In addition, an Chinese multi-modal depressed patients database will be recored, and a multi-modal depression recognition (audio, video, physiological signals, text, etc.) network will be pursued to identify manifestations from depressed patients among different cultures and races.

Acknowledgment

This work is supported by the Shaanxi Provincial Office of Education Emergency Research Fund for Public Health Security (grant 20JG030), the Shaanxi Higher Education Association Fund for the Prevention and Control of Novel Coronavirus Pneumonia (grant XGH20201), the Shaanxi Provincial Public Scientific Quality Promotion Fund for Emergency Popularization of COVID-19 (grant 2020PSL(Y)040).

References

[1] C. Mathers, D. M. Fat, J. T. Boerma, The global burden of disease: 2004 update, World Health Organization, 2008.

[2] L. Wen, X. Li, G. Guo, Y. Zhu, Automated depression diagnosis based on facial dynamic analysis and sparse coding, IEEE Transactions on Information Forensics and Security 10 (7) (2015) 1432–1441.

[3] L. He, D. Jiang, H. Sahli, Multimodal depression recognition with dynamic visual and audio cues, in: International Conference on Affective Computing and Intelligent Interaction, 2015, pp. 260–266.

[4] A. Dhall, R. Goecke, A temporally piece-wise fisher vector approach for depression analysis, in: Affective Computing & Intelligent Interfaces, 2015, pp. 255–259.

[5] S. Song, L. Shen, M. Valstar, Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 158–165.

[6] X. Zhou, K. Jin, Y. Shang, G. Guo, Visually interpretable representation learning for depression recognition from facial images, IEEE Transactions on Affective Computing (2018) 1–1.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1–9.

[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[10] D. Yi, Z. Lei, S. Liao, S. Z. Li, Learning face representation from scratch, arXiv preprint arXiv:1411.7923.

[11] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, AVEC 2013: the continuous audio/visual emotion and depression recognition challenge, in: Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge, ACM, 2013, pp. 3–10.

[12] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, AVEC 2014: 3D dimensional affect and depression recognition challenge, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 3–10.

[13] M. Al Jazaery, G. Guo, Video-based depression level analysis by encoding deep spatiotemporal features, IEEE Transactions on Affective Computing.

[14] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 445–450.

[15] S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2012) 221–231.

[16] A. T. Beck, R. A. Steer, R. Ball, W. F. Ranieri, Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients, Journal of Personality Assessment 67 (3) (1996) 588–97.

[17] J. M. Girard, J. F. Cohn, M. H. Mahoor, Nonverbal social withdrawal in depression: evidence from manual and automatic analyses, Image and Vision Computing 32 (10) (2014) 641–647.

[18] L. He, D. Jiang, H. Sahli, Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding, IEEE Transactions on Multimedia 21 (6) (2018) 1476–1486.

[19] Y. Zhu, Y. Shang, Z. Shao, G. Guo, Automated depression diagnosis based on deep networks to encode facial appearance and dynamics, IEEE Transactions on Affective Computing 9 (4) (2017) 578–584.

[20] A. Jan, H. Meng, Y. F. B. A. Gaus, F. Zhang, Artificial intelligent system for automatic depression level analysis through visual and vocal expressions, IEEE Transactions on Cognitive and Developmental Systems 10 (3) (2017) 668–680.

[21] W. C. de Melo, E. Granger, A. Hadid, Combining global and local convolutional 3d networks for detecting depression from facial expressions, FG, 2019.

[22] S. Song, S. Jaiswal, L. Shen, M. Valstar, Spectral representation of behaviour primitives for depression analysis, IEEE Transactions on Affective Computing (2020) 1–1.

[23] J. B. J. Md Azher Uddin, Y.-K. Lee, Depression level prediction using deep spatiotemporal features and multilayer bi-ltsm, IEEE Transactions on Affective Computing (2020) 1–1.

[24] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–10.

[25] D. A. AL CHANTI, A. Caplier, Deep learning for spatio-temporal modeling of dynamic spontaneous emotions, IEEE Transactions on Affective Computing.

[26] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2983–2991.

[27] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using cnn with attention mechanism, IEEE Transactions on Image Processing 28 (5) (2018) 2439–2450.

[28] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, D. N. Metaxas, Learning active facial patches for expression analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 2562–2569.

[29] C. Yan, B. Gong, Y. Wei, Y. Gao, Deep multi-view enhancement hashing for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 1–1.

[30] C. Yan, B. Shao, H. Zhao, R. Ning, Y. Zhang, F. Xu, 3d room layout estimation from a single rgb image, IEEE Transactions on Multimedia (2020) 1–1.

[31] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, IEEE, 2006, pp. 2169–2178.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.

[33] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

[34] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, J. Epps, Diagnosis of depression by behavioural signals: A multimodal approach, in: Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge, ACM, ACM, Barcelona, Spain, 2013, pp. 11–20.

[35] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, Y. Wang, Depression recognition based on dynamic facial and vocal expression features using partial least square regression, in: Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge, ACM, ACM, Barcelona, Spain, 2013, pp. 21–30.

[36] W. C. de Melo, E. Granger, A. Hadid, Depression detection based on deep distribution learning, ICIP, 2019.

[37] M. Sidorov, W. Minker, Emotion recognition and depression diagnosis by acoustic and visual features: A multimodal approach, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, ACM, Orlando, Florida, USA, 2014, pp. 81–86.

[38] A. Jan, H. Meng, Y. F. A. Gaus, F. Zhang, S. Turabzadeh, Automatic depression scale prediction using facial expression dynamics and regression, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, ACM, Orlando, Florida, USA, 2014, pp. 73–80.

[39] H. Kaya, F. Çilli, A. A. Salah, Ensemble cca for continuous emotion prediction, in: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, ACM, 2014, pp. 19–26.