# Depression Level Assessment based on 3D CNN and Facial Expression Videos

Junlong Gao [1, a], Yucheng Wei [1, b]

[1] Department of Automation, Faculty of Information Technology, Beijing University of Technology, Beijing; 100124, China;

[a] 744145154@qq.com, [b] 1170411311@qq.com

**Abstract.** Depression has a severe impact on people's daily lives and work, and it may even lead to suicide. Computer visionbased methods are promising for providing more effective and objective assistance in the clinical diagnosis of depression. In this article, to compare the performance of different 3D convolutional neural networks in assessing depression levels, we tested 3D VGGNet18, 3D GoogleNet, 3DEfficientNetB7, and 3D MobileNetV3 networks based on the AVEC2013 and AVEC2014 datasets. Experimental results showed that the 3D MobileNetV3 network achieved the best evaluation results, with MAE=7.35 and RMSE=9.16 on the AVEC2013 dataset, and MAE=7.19 and RMSE=9.08 on the AVEC2014 dataset. Compared with other existing methods, 3D ResNet18 demonstrated excellent performance.

**Keywords:** Deep Learning; Depression; 3D CNN.

## 1. Introduction

Depression stands as a major contributor to psychological health issues across diverse age groups. Statistics indicate a lifetime prevalence of approximately 10-20% in females and 5-12% in males. This mental disorder profoundly affects individuals' cognition, behavior, emotions, and work capabilities, often leading to feelings of sadness, helplessness, anxiety, despair, worry, anger, or restlessness. In severe cases, depression may even culminate in suicidal tendencies.

Fortunately, effective treatments for depression, including medication, psychological counseling, and other clinical approaches, exist. However, current diagnostic methods for depression heavily rely on comprehensive assessments by experienced professionals. Nevertheless, these diagnostic approaches are constrained by subjective factors and the lack of precise measurement standards, potentially introducing biases. With the increasing prevalence of depression cases, the need for accurate diagnosis becomes increasingly urgent.

Previous research has demonstrated that individuals suffering from depression exhibit distinct speech patterns when compared to those who are healthy, leading to the development of some methods that utilize audio cues for automated depression diagnosis. Additionally, nonverbal cues, such as gestures and facial expressions, are also believed to reflect the severity of depression. In human visual communication, more than half of nonverbal behaviors involve the facial area. Therefore, this study will focus on methods for assessing depression levels based on facial expression signals.The currently commonly used method is to utilize the AVEC2013 and AVEC2014 datasets.

Meng et al.[2] employed Motion History Histograms (MHH) as a framework for image representation. This approach involved deriving five MHH-based representations from each facial image, followed by the extraction of Edge Orientation Histograms (EOH) and Local Binary Patterns (LBP) features. On the other hand, Jan et al.[3] proposed the utilization of Local Binary Patterns (LBP), Edge Orientation Histograms (EOH), and Local Phase Quantization (LPQ) as feature descriptors to extract features from images. Subsequently, they introduced a one-dimensional Motion History Image (MHH) technique to capture changes in each component across a sequence of feature vectors. For prediction, they employed Partial Least Squares Regression.

Recently, an increasing number of researchers have been using deep learning techniques to automatically extract features from images. Zhu et al. [4] proposed a two-stream convolutional neural network that captures both static facial appearance features and dynamic features from images The first stream inputs the facial region, while the second stream inputs the facial flow. Finally, two fully connected layers are used to fuse these features and make predictions. Zhou et al. [5] proposed DepressNet to extract facial features for predicting BDI-II scores, and based on this, they further introduced MR DepressNet. They divided the facial region into different areas and input each area into DepressNet to predict the score for that specific area.Take the average of all prediction scores for depression level as the final score.

In this research, we opted for several classic 2D Convolutional Neural Networks (CNNs), including VGGNet, GooleNet, EfficientNet and MobileNetV3. We modified them into their corresponding 3D CNN forms and utilized the AVEC2013 and AVEC2014 datasets to evaluate depression levels, exploring the performance of different 3D CNNs in depression level assessment. Due to interference from lighting and background in the original video recordings, the collected video quality was inadequate. Therefore, we employed data augmentation methods to enhance the performance of depression level assessment.

## 2. Depression Level Assessment Model

### 2.1 3D CNN

Video signals are three-dimensional data integrating spatial and temporal information. Utilizing 2D CNNs only allows modeling spatial information, neglecting the role of temporal information. Figure 1 illustrates the distinction between 2D and 3D convolutions. 3D convolutional kernels operate in three dimensions (depth, height, and width), enabling the simultaneous extraction of spatial and temporal features from input data.

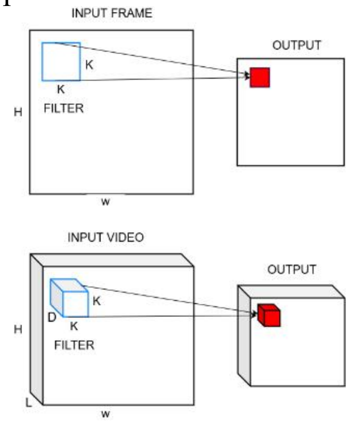

Figure 1. Iillustrates the difference between 2D CNN and 3D CNN convolutions. In 2D CNN (top), a convolution kernel of size is used to convolve the image, generating a feature map. In contrast, 3D CNN (bottom) performs 3D convolutions on an image sequence using a convolution kernel of size , resulting in a volumetric feature.

### 2.2 3D Convolutional Neural Network

VGGNet, GoogleNet, EfficientNet, and MobileNetV3 are currently commonly used networks in image recognition.

VGGNet:The core idea of VGGNet is to replace larger convolutional layers with multiple smaller convolutional layers. For example, stacking three 3x3 convolutional layers is equivalent to a 7x7 convolutional layer in terms of receptive field. This not only reduces parameters but also enhances the CNN's ability to learn features by introducing more non-linear transformations.

GoogleNet:Compared to the early AlexNet model, GoogleNet has a deeper network structure with a total of 87 layers. Despite its increased complexity, GoogleNet has fewer parameters, mainly due to the introduction of the Inception module. This module applies multiple convolutional kernel sizes and pooling operations in parallel, then concatenates their outputs. This structure is beneficial for extracting features at different scales while reducing the overall number of model parameters.

EfficientNet:EfficientNet is to adopt a mixed-dimension model scaling method that adjusts the size of the model by scaling the three dimensions of depth, width, and resolution. EfficientNet is able to improve model performance without significantly increasing computational complexity.

MobileNetV3:MobileNetV3 is an upgraded version of MobileNetV1 and MobileNetV2, incorporating the previous depthwise separable convolutions as well as residual structures. With smaller parameters and computational costs, MobileNetV3 is well-suited for use in scenarios with limited storage space and exhibits excellent performance.

In this study, we experimented with the 3D structures of the mentioned models (3D VGGNet16, 3D GoogleNet, 3D EfficientNetB7, and 3D MobileNetV3) on AVEC2013 and AVEC2014 datasets. We replaced the 2D structures with their corresponding 3D structures, maintaining the overall network depth, with the input being three-dimensional video data.

## 3. Datasets

### 3.1 AVEC2013 Depression Dataset

AVEC2013 originated from the Audio/Visual Emotion Challenge in 2013 [6]. In this competition, 82 subjects participated, and each subject was interviewed through human-computer interaction. During the interview, data were recorded using a camera and a microphone. The dataset comprises three parts: training, development, and test sets. Each part contains 50 videos, averaging about 25 minutes in length.

### 3.2 AVEC2013 Depression Dataset

AVEC2014 originated from the Audio/Visual Emotion Challenge in 2014 [7]. The participants in this competition were involved in only two tasks: the Freeform task and the Northwind task. In the Freeform task, the participants answered questions similar to "Discuss a sad childhood memory." In the Northwind task, the participants read aloud an excerpt from a fable. The recorded videos were also divided into three parts. Each part included both tasks, with each task containing 50 videos. The duration of these videos ranged from 6 seconds to 248 seconds.

### 3.3 Dataset Preprocessing

3.3.1 Subsampling

For AVEC2013 and AVEC2014, considering the redundancy in video frames, we initially reduced the total number of frames per video input through frame subsampling. Given the varying durations of videos in each dataset, we sampled every 100 frames from AVEC2013 and every 10 frames from AVEC2014, taking every 16 frames as input for the model. There is an overlap of 8 frames between adjacent sequences of 16 frames.

3.3.2 Face Detection and Alignment:

The position of a person's head can unconsciously shift during video recording. We utilized the Dlib face detection tool to detect 64 facial landmarks, including eyes, mouth, and nose, and cropped the facial images based on these landmarks. All images were aligned with the first frame, and the images were resized.

3.3.3 Image Enhancement

In AVEC2013 and AVEC2014, the videos in the datasets were subject to interference from external factors such as lighting and the participants' head poses. Therefore, we employed the Adaptive Histogram Equalization (AHE) method for image enhancement.

### 3.4 Dataset Preprocessing

We merged the original training sets, enhanced training sets, original development sets, and enhanced development sets of AVEC2013 and AVEC2014 into a new training set for model training. The model's performance was assessed utilizing the enhanced test set.

### 3.5 Experiment setting

The network model was built using the TensorFlow toolkit. A stochastic gradient descent (SGD) optimizer was adopted. The batch size was set to 10, and the learning rate was used $7 \times 10^{-7}$. We conducted experiments on a Tesla V100 GPU.

We use MAE and RMSE to measure model performance:

$$MAE = \frac{1}{M}\sum_{i=0}^{M-1} |x_i - \hat{x}_i| \#(1)$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=0}^{M-1} (x_i - \hat{x})^2 \#(2)}$$

Where $x_i$ represents the true value, $\hat{x}_i$ represents the predicted value, and M represents the total number of samples.

## 4. Experimental Results

### 4.1 Results of experiments conducted on various models

We conducted experiments using 3D VGGNet16, 3D Goolnet, 3D EfficientNetB7, and 3D MobileNetV3 models on AVEC2013 and AVEC2014. The findings are summarized in Table 1 and Table 2. One can notice that the 3D MobileNetV3 network achieved the best results. Additionally, we compared the experimental results before and after image enhancement, and it was found that the enhancement of images can considerably enhance the accuracy of depression level evaluation.

Table 1. Experimental Results Before Image Enhancement

| Metheds | AVEC2013 | | AVEC2014 | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| 3D VGG16Net | 8.55 | 10.93 | 8.21 | 10.81 |
| 3D GooleNet | 8.11 | 10.47 | 7.96 | 10.48 |
| 3D EfficientNetB7 | 7.91 | 10.66 | 7.59 | 10.13 |
| 3D MobileNetV3 | 7.65 | 10.12 | 7.26 | 9.31 |

Table 2. Experimental Results After Image Enhancement

| Metheds | AVEC2013 | | AVEC2014 | |
|---|---|---|---|---|
| | MAE | RMSE | MAE | RMSE |
| 3D VGG16Net | 8.36 | 10.96 | 8.09 | 10.72 |
| 3D GooleNet | 7.94 | 10.71 | 7.78 | 10.65 |
| 3D EfficientNetB7 | 7.76 | 10.15 | 7.51 | 10.19 |
| 3D MobileNetV3 | 7.35 | 9.16 | 7.19 | 9.08 |

### 4.2 Comparison with Other Depression Level Assessment Methods

We compared 3D MobileNetV3 with other existing methods on the AVEC2013 and AVEC2014 datasets, as shown in Table 3 and Table 4. Deep learning outperforms manual feature extraction methods because it can automatically extract features. The experiments demonstrate that 3D convolutional neural networks can enhance the accuracy of depression level assessment.

Table 3. Comparison with Other Existing Methods on the AVEC2013 Dataset

| Methods | MAE | RMSE |
|---|---|---|
| Baseline[6] | 10.88 | 13.61 |
| Kächele et al. [8] | 8.97 | 10.82 |
| Wen et al. [9] | 8.22 | 10.27 |
| Kaya et al.[10] | 7.86 | 9.72 |
| Zhu et al. [4] | 7.58 | 9.82 |
| Mohamad et al. [11] | 7.37 | 9.28 |
| 3D MobileNetV3 | 7.35 | 9.16 |

Table 4. Comparison with Other Existing Methods on the AVEC2014 Dataset

| Methods | MAE | RMSE |
|---|---|---|
| Baseline[7] | 8.86 | 10.86 |
| InaoeBuap [12] | 8.46 | 9.84 |
| Brunel [13] | 8.44 | 10.50 |
| BU-CMPE [14] | 8.20 | 10.27 |
| Zhu et al. [4] | 7.47 | 9.55 |
| Mohamad et al. [11] | 7.22 | 9.20 |
| 3D MobileNetV3 | 7.19 | 9.08 |

## 5. Conclusion

In this paper, we assessed the performance of various 3D neural networks, including 3D VGGNet16, 3D Goolnet, 3D EfficientNetB7, and 3D MobileNetV3, for depression level evaluation using the AVEC2013 and AVEC2014 datasets. Our experimental findings indicated that the 3D MobileNetV3 network exhibited superior performance. Furthermore, comparative experiments revealed that adaptive contrast adjustment can effectively enhance the accuracy of depression level assessment. In our future work, we aim to incorporate audio, text, and other types of data to conduct a more comprehensive analysis of depression.

## References

[1] Beck A T, Steer R A, Ball R, et al. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients[J]. Journal of personality assessment, 1996, 67(3): 588-597.

[2] H. Meng and N. Pears, “Descriptive temporal template features for visual motion recognition,” Pattern Recognition Letters, vol. 30, no. 12, pp.1049–1058, 2009.

[3] Jan A, Meng H, Gaus Y F A, et al. Automatic depression scale prediction using facial expression dynamics and regression[C]//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. 2014: 73-80.

[4] 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.

[5] Zhou X, Jin K, Shang Y, et al. Visually interpretable representation learning for depression recognition from facial images[J]. IEEE transactions on affective computing, 2018, 11(3): 542-552.

[6] M. V alstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia,S. Schnieder, R. Cowie, and 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.

[7] M. V alstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski,R. Cowie, and 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.

[8] M. V alstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski,R. Cowie, and 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.

[9] Wen L, Li X, Guo G, et al. Automated depression diagnosis based on facial dynamic analysis and sparse coding[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(7): 1432-1441.

[10] Kaya H, Salah A A. Eyes whisper depression: A CCA based multimodal approach[C]//Proceedings of the 22nd ACM international conference on Multimedia. 2014: 961-964.

[11] Al Jazaery M, Guo G. Video-based depression level analysis by encoding deep spatiotemporal features[J]. IEEE Transactions on Affective Computing, 2018, 12(1): 262-268.

[12] Pérez Espinosa H, Escalante H J, Villaseñor-Pineda L, et al. Fusing affective dimensions and audio-visual features from segmented video for depression recognition: INAOE-BUAP's participation at AVEC'14 challenge[C]//Proceedings of the 4th international workshop on audio/visual emotion challenge. 2014: 49-55.

[13] Jan A, Meng H, Gaus Y F A, et al. Automatic depression scale prediction using facial expression dynamics and regression[C]//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. 2014: 73-80.

[14] Kaya H, Çilli F, Salah A A. Ensemble CCA for continuous emotion prediction[C]//Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. 2014: 19-26.