90160_Auto_Edited .docx

WORD COUNT	3996	TIME SUBMITTED	11-JAN-2024 11:32PM
		PAPER ID	106086340

Name of Journal: *World Journal of Psychiatry*

Manuscript NO: 90160

Manuscript Type: MINIREVIEWS

Automatic Recognition of Depression Based on Audio and Video: A Review

Han M *et al*. The research progress of ADE.

Mengmeng Han, Xingyun Li, Xinyu Yi, Yunshao Zheng, Weili Xia, Yafei Liu, Qingxiang Wang

Abstract

Depression is a common mental health disorder. With current depression detection methods, specialized physicians often engage in conversations and physiological examinations based on standardized scales as auxiliary measures for depression assessment. Non-biological markers— typically classified as verbal or non-verbal and deemed crucial evaluation criteria for depression—have not been effectively utilized. Specialized physicians usually require extensive training and experience to capture changes in these features. Advancements in deep learning technology have provided technical support for capturing non-biological markers. Several researchers have proposed automatic depression estimation (ADE) systems based on sounds and videos to assist physicians in capturing these features and conducting depression screening. This article summarizes commonly used public datasets and recent research on audio- and video-based ADE based on three perspectives: datasets, deficiencies in existing research, and future development directions.

Key Words: Depression Recognition; Deep Learning; Automatic Depression Estimation (ADE) System; Audio Processing; Image Processing; Feature Fusion; Future Development

Han M, Li X, Yi X, Zheng Y, Xia W, Liu Y, Wang Q. Automatic Recognition of Depression Based on Audio and Video: A Review. *World J Psychiatry* 2023; In press

Core Tip: The automatic recognition of depression based on deep learning has gradually become a research hotspot. Researchers have proposed Automatic Depression Estimation (ADE) systems utilizing sound and video data to assist physicians in screening for depression. This article provides an overview of the latest research on ADE systems, focusing on sound and video datasets, current research challenges, and future directions.

INTRODUCTION

With societal developments, the diagnosis and treatment of depression have become increasingly crucial. Depression is a prevalent psychological disorder characterized by symptoms such as low mood, diminished appetite, and insomnia in affected individuals[1]. Patients with severe depression may also exhibit a tendency towards suicide. In the field of medicine, researchers aspire to conduct comprehensive investigations of depression from both biological and non-biological perspectives. Li *et al*[2] summarized biological markers, revealing associations between depression and indicators, such as gamma-glutamyl transferase, glucose, triglycerides, albumin, and total bilirubin. Non-biological markers can be broadly categorized into verbal and non-verbal features. Verbal features typically pertain to a subject's intonation, speech rate, and emotional expressions in speech extracted from audio recordings. Early studies by Cannizzaro *et al*[3] and Leff *et al*[4] identified differences in the speech of individuals with psychiatric disorders compared to the general population. Non-verbal features typically refer to the facial expressions and body movements commonly embedded in video files. The Facial Action Coding System (FACS)[5], a frequently employed tool for facial expression analysis, decomposes facial muscles into multiple action units (AUs) with corresponding numerical identifiers. For instance, AU1 and AU2 represent inner brow raise and outer brow raise, respectively. A graphical representation of the AU can be accessed through the link indicated in the footnote [1]. While Girard *et al*[6] found differences in AU 10, 12, 14, and 15 between individuals with depression and the general population, a unified research framework for bodily changes is yet to be established, with the core challenge lying in quantifying alterations in body movements. Joshi *et al* demonstrated the potential of studying body movements for ADE using a method based on space-time interest points and a bag of words to analyze patients' upper-body movements.

During clinical assessments, specialized physicians detect and treat depression based on diagnostic criteria manuals issued by the relevant organizations. For instance, the World Health Organization (WHO) released the 11th revision of the International Classification of Diseases (ICD-11) in 2022, providing detailed classifications of various mental disorders. The American Psychiatric Association published the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-4)[7], in 1994, and its updated version, DSM-5[8], in 2013. In 2001, China released the Chinese Classification and Diagnostic Criteria of Mental Disorders, Third Edition (CCMD-3). Guided by diagnostic manuals, specialized physicians assessed the severity of depression in the participants based on the scores obtained from these scales. Rating scales are typically categorized into self-report and observer-report scales. The patient health questionnaire (PHQ)[9] is a lightweight self-report scale, whereas the Hamilton depression rating scale (HAMD)[10] is a common observer-report scale. Observer-report scales require specialized physicians to interview patients and score the details based on the scale. Completing an interview based on the HAMD scale typically takes 15–20 minutes.

In addition to detecting clinical depression based on rating scales, biological markers have been employed to assist with the assessment. Physicians use biochemical indicators extracted through techniques, such as blood tests, to aid their judgment. With the advancements in detection technologies, biological markers can be quantitatively measured, allowing specialized physicians to directly refer to numerical values to determine the clinical significance of a test. However, non-biological markers, which are crucial features of depression, have not been extensively utilized, attributed to several factors. First, changes in non-biological markers, such as facial expressions and intonation, are often subtle. Specialized physicians require extensive training and accumulated experience to capture these changes; such training is typically time-consuming and inefficient. Second, unlike biological markers, systematic patterns of change in non-biological markers depend on their ability to capture spatial and temporal information, a challenging task for early computer technologies. The development of deep-learning technology and the computational capabilities of computers provide an opportunity to address these challenges. Deep-learning, with its robust capability of capturing temporal and spatial information, offers new avenues for constructing assistive systems. Automatic depression estimation (ADE) has become a significant research direction in the field of computational medicine, resulting in several ADE methods being proposed.

A complete ADE study typically comprises three steps. The first step involves data collection, categorized based on the free or need for specific emotional stimulus experiments. The former typically utilizes devices such as cameras and microphones to capture audio-visual information of subjects during medical consultations or in natural states. The latter requires the design of specific emotional paradigms, followed by recording subjects' audio-visual information under emotional stimuli. The second step involves constructing deep-learning models for ADE. In this phase, researchers designed different deep-learning architectures based on data characteristics to capture information for ADE. Finally, the model undergoes training and testing for ADE to essentially perform two tasks: classification, i.e., distinguishing whether the individual is a patient with depression or further categorizing the severity (non-depressed, mild, moderate, and severe), and scoring tasks, i.e., predicting the assessment scale scores of the subjects. Depending on the task, researchers choose different evaluation metrics to train and test the effectiveness of the model.

The initial ADE typically requires manual feature extraction and the application of machine learning methods such as decision trees and support vector machines for feature classification. Peng *et al*[11] initially constructed a sentiment lexicon, counted word frequencies, and then input these features into a support vector machine for ADE. Alghowinem *et al*[12] first used the openSMILE tool to extract audio features and then employed machine learning methods for ADE. Wen *et al*[13] extracted dynamic feature descriptors from facial region sub-volumes and used sparse coding to implicitly organize the extracted feature descriptors for depression diagnosis. With the development of deep learning and computational capabilities, deep models can perform feature extraction from complex data, eliminating manual feature extraction. Notably, owing to the specificity of audio information, certain manual feature extraction steps still exist. Therefore, a series of deep learning-based ADE methods, such as [17-39], have been proposed. In this review, we focus primarily on recent ADE methods based on deep learning approaches. We first introduce commonly used publicly available ADE datasets and then provide an overview and summary of recent outstanding audio-visual ADE models. All articles are summarized in Table 1. Finally, we summarize the existing challenges and future directions of ADE.

DATASETS

While data form the foundation of ADE research, owing to the inherent challenges in collecting depression data, such as strong privacy concerns, lengthy collection periods, and limited data volumes, obtaining subject authorization for public sharing is difficult, resulting in a scarcity of publicly available audio-visual datasets. Commonly utilized public datasets primarily originate from audio-visual emotion recognition challenges, specifically the AVEC2013[14], AVEC2014[15], and DAIC-WOZ[16] datasets.

AVEC2013: The AVEC2013 dataset was released as part of the third Audio-Visual Emotion Recognition Challenge (AVEC2013). This dataset comprises 340 video segments collected from 292 participants. AVEC2013 required participants to perform tasks such as vowel phonation, reading, recounting memories, and narrating a story based on a picture with their audio-visual information recorded. The Beck Depression Inventory (BDI) scores served as labels for AVEC 2013.

AVEC2014: The AVEC2014 dataset was released as part of the fourth Audio-Visual Emotion Recognition Challenge (AVEC2014). This dataset comprises 150 audio-video data segments involving a total of 84 subjects. As a subset of AVEC2013, AVEC2014 required each participant to complete two tasks, Northwind and Freeform, which involved reading excerpts from articles and answering specific questions. Similar to AVEC2013, AVEC2014 also utilizes BDI scores as data labels.

Distress Analysis Interview Corpus/Wizard-of-Oz set (DAIC-WOZ): This dataset encompasses the audio-visual information of subjects collected through various interview formats, with each data type being independent. The video information, which included a maximum of 263 audio-visual data points, was based on facial features (e.g., annotated directions, facial key points, and AU features) after conversion.

AUDIO-BASED DEPRESSION ESTIMATION

Audio-based methods are crucial for ADE. In this process, participants often combine manual features with deep features for ADE. Manual features typically include time- and frequency-domains. Deep features are typically obtained from spectrograms using deep-learning models. These spectrograms often represent the waveform, spectrogram, Mel spectrogram, or processed data of raw audio graphically.

He *et al*[17] combined manually extracted audio features with deep-learning features for ADE. They divided the model into two parts. The first part employed a deep network to extract deep features from spectrograms and raw speech waveforms. The other part extracts median robust extended local binary patterns (MRELBP) from spectrograms and low-level descriptors from raw speech. Finally, these features were fused using a fusion model to make the final decision. This approach achieved RMSE and MAE values of 10.001 and 8.201 on the AVEC2013 dataset and 9.999 and 8.191 on the AVEC2014 dataset. Zuo *et al*[18] recognized the potential performance decline associated with limited audio data. With a smaller dataset, capturing the patterns of depressive expressions becomes challenging, and deep models tend to learn audio features specific to individual subjects, leading to overfitting. To address this issue, they proposed a speaker-invariant depression detector (SIDD), which achieved an F1 score of 0.601 on the DAIC-WOZ dataset. Du *et al*[19] incorporated patients' vocal tract changes into conventional speech perceptual features and developed a machine speech chain model for depression recognition (MSCDR) for ADE. The MSCDR extracts speech features from both generation and perception aspects and uses recurrent neural networks (RNN) to extract time-domain features for depression detection. The MSCDR achieved accuracy and F1 scores of 0.771 and 0.746, respectively, on the DAIC-WOZ dataset.

Yang *et al*[20] observed that many ADE models based on manually designed features lack good interpretability, with features not fully utilized. Therefore, the DALF was proposed. Learnable filters in DALF can more effectively decompose audio signals and retain effective features. Analyzing the automatically learned filters allows for a deeper understanding of the focus areas of the model. This method achieved an F1 score of 0.784 on the DAIC-WOZ dataset. Han *et al*[21] introduced a spatial-temporal feature network (STFN) to capture audio features. The STFN initially captured the deep features of audio information and then used a novel mechanism called hierarchical contrastive predictive coding (HCPC) loss, replacing the commonly used RNN to capture temporal information. This approach reduces the parameter count of the model, making it more trainable. As such, the STFN achieved accuracy, RMSE, and MAE values of 0.780, 6.36, and 5.38, respectively, on the DAIC-WOZ dataset. Chen *et al*[22] focused on integrating the Transformer architecture with audio features. Their proposed model, SpeechFormer++, utilized prior knowledge to guide feature extraction, achieving an accuracy of 0.733% on the DAIC-WOZ dataset. Mao *et al*[23] recognized that text features in audio are also important for capturing the patterns of depressive expressions. Consequently, they proposed an attention-based fused representation of text and speech features. This approach initially inputs text information and low-level features of raw speech into an encoder for encoding and subsequently employs the encoded features for depression detection, achieving an F1 score of 0.958 in a five-class classification task using the DAIC-WOZ dataset.

Overall, the design of ADE models based on audio relies on the initial feature selection. Because audio information cannot be utilized directly by deep models, it is typically transformed before being extracted by deep models. These transformations are diverse, including directly using the raw waveform, applying Fourier transform or Fast Fourier Transform to transform the time–frequency domain information, converting audio into Mel spectrograms, and directly extracting audio features such as frame intensity, frame energy, and fundamental frequency. Diverse feature selection methods provide various possibilities for ADE, leading to discussions regarding which audio representation is more beneficial for ADE. The construction of the model must be aligned with the selected features for an effective feature extraction. Given that depression datasets are often small, methods to limit the learning of individual features by the model, as demonstrated by Zuo *et al*[18], should be carefully considered.

VIDEO-BASED DEPRESSION ESTIMATION

Video information often preserves changes in participants' facial expressions during exposure to stimulus paradigms. Facial expressions include both facial and bodily expressions. In medical research, video-based ADE models typically incorporate various attention mechanisms to enhance local facial features. He *et al*[24] proposed an ADE framework called the deep local global attention convolutional neural network (DLGA-CNN). In the DLGA-CNN, multiple attention mechanisms are introduced and utilized for extracting multiscale local and global features. Finally, these multiscale features are fused and employed for depression detection. The DLGA-CNN achieved RMSE and MAE values of 8.39 and 6.59 on AVEC2013 and 8.30 and 6.51 on AVEC2014. He *et al*[25] also recognized the presence of annotation noise in depression datasets, which could negatively affect feature extraction and result in suboptimal ADE performance. Therefore, they proposed a self-adaptation network (SAN) to relabel erroneous annotations in the datasets. SAN achieved RMSE and MAE values of 9.37 and 7.02 on AVEC2013 and 9.24 and 6.95 on AVEC2014.

Zhao[26] acknowledged the significance of local and global information and proposed an ADE architecture based on facial images. To enhance the quality of facial images, the architecture initially utilizes the Gamma Correction[27] and DeblurGAN-v2[28] algorithms to balance brightness and contrast and improve image clarity. The architecture employs ConvFFN[29] as the main framework and designs the Hi-Lo attention module to enhance the features in different facial regions. Ultimately, this method achieved RMSE and MAE values of 7.36 and 5.97 on the AVEC2013 dataset and 7.23 and 5.85 on the AVEC2014 dataset. Liu *et al*[30] introduced another approach, PRA-Net, for feature extraction from facial regions for ADE. PRA-Net initially segments the extracted facial feature maps by region; these segmented regions are fed into a self-attention mechanism to capture interregional correlations. The classifier merges the regional feature maps with weights for the final decision. PRA-Net achieved RMSE and MAE values of 7.59 and 6.08 on the AVEC2013 dataset and 7.98 and 6.04 on the AVEC2014 dataset.

In addition to extracting features from the entire face, Yuan *et al*[31] explored the use of gaze features for ADE. They employed a fully connected network to extract gaze features from the participants and achieved an accuracy value of 0.831. Subsequently, Zhao *et al*[32] designed an attention-based architecture, EnSA, for ADE that achieved an accuracy of 0.955.

In addition to using facial expressions for ADE, utilizing body expressions is also an important approach. Yu *et al*[33] initially captured the participants' body skeleton change sequences using Kinect. Subsequently, they constructed a spatial attention-dilated TCN (SATCN) based on an improved temporal convolutional network (TCN). SATCN achieved a maximum accuracy of 0.758 for binary classification tasks and a maximum accuracy value of 0.643 for multiclass datasets. Similarly, Zhao *et al*[34] employed body skeletal information for ADE. They observed differences in reaction times between the case and control groups for specific tasks. Consequently, they used the reaction time as prior knowledge along with skeletal information and input them into a Transformer for ADE. This approach achieved an accuracy value of 0.729. Compared to the abundance of facial-based ADE models, the number of models based on body expressions is relatively limited, warranting further research and exploration.

Unlike audio information, the advancement of convolutional networks enables the direct utilization of image information. Consequently, the construction of end-to-end ADE models has become mainstream in recent years. The inputs for these models do not require complex preprocessing and typically involve region cropping and lighting balancing. Extracting local information has become a crucial aspect of model construction and has emerged as a primary research direction for ADE based on video information.

FUSION OF AUDIO- AND VISUAL-BASED DEPRESSION ESTIMATION

In addition to using unimodal information for depression prediction, depression-detection models that jointly utilize multiple modalities are being developed. Various methods of complementing information enhance the accuracy of multi-modal models compared to unimodal models, with the combination of audio and visual information a commonly used approach.

Yang *et al*[35] designed uncertainty-aware label contrastive and distribution learning (ULCDL) to integrate facial, audio, and text information for ADE. ULCDL introduces a contrastive learning framework into ADE to enhance a model's learning capability, achieving an accuracy value of 0.830 and F1 score of 0.900 on the DAIC-WOZ dataset. Niu *et al*[36] combined facial sequences with audio spectrograms to detect ADE. Leveraging the characteristics of both features, they proposed spatiotemporal attention (STA) and multi-modal attention feature fusion (MAFF) networks to enhance and obtain cross-modal attention for the two features. This architecture achieved RMSE and MAE values of 8.16 and 6.14 on AVEC2013 and 7.03 and 5.21 on AVEC2014. Shao *et al*[37] observed that different features from the same data can be complementary. They combined the participants' RGB images of the body and body skeleton images for the

ADE, achieving an accuracy value of 0.854 on a dataset comprising 200 participants. Zhou *et al*[38] approached ADE from the perspective of video blogs. Their proposed time-aware attention-based multi-modal fusion depression detection network (TAMFN) extracts and fuses multi-modal information from three aspects: global features, inter-modal correlations, and temporal changes. TAMFN obtained an F1 score of 0.75 on the D-Vlog[39] dataset. Uddin *et al*[40] initially segmented audio and video into equally sized segments before using volume local directional structural patterns (VLDSPs) and temporal attention pooling (TAP) to encode facial and audio information to obtain the importance of each video and audio segment. The next step involved formatting video and audio segments. Finally, multi-modal factorized bilinear pooling (MFB) was employed to fuse the features and make decisions. This method achieved RMSE and MAE values of 6.83 and 5.38 on AVEC2013 and 6.16 and 5.03 on AVEC2014.

Multimodality is a new approach to ADE. Multimodal information mimics the patterns of diagnosis and treatment from multiple perspectives in clinical examinations. The most crucial aspect of multimodal information is the exploration and integration of hidden relationships among the various types of information. Initially, feature and decision fusions were the primary methods for combining features. However, these two approaches are simple and do not consider deep feature integration. With further research, ADE will demand multiscale and deep fusion of multimodal features. Cross-modal fusion methods are no longer limited to feature and decision fusions. When constructing new fusion methods, identifying relationships between different types of information and methods to capture these relationships become crucial.

DISCUSSION

Facial information has been favored by most researchers for ADE methods based on video information. Studies such as [6] subdivided faces into multiple AUs for investigation. Inspired by these studies, researchers recognized the importance of local facial information. A series of attention mechanisms were proposed and employed to facilitate the model's focus on local information. Although researches[33,34] has explored aspects such as gait and body movements in individuals with depression, compared with the excessive attention paid to ADE methods based on facial information, methods based on body expressions for ADE appear to be relatively scarce. Notably, databases analyzing the body movements of individuals with depression are often not publicly accessible. In addition, publicly available depression databases rarely contain body information. The lack of visibility and the difficulty in data collection are significant reasons for the limited development of ADE methods based on body expressions. Despite these challenges, we believe that this is an important research direction as facial expressions. We hope to develop more ADE methods based on the proposed body changes. Research related to human motion recognition, keypoint capture, and skeleton tracking may serve as valuable references for constructing ADE models based on body expressions.

For audio-based ADE methods, current approaches primarily involve combining handcrafted features or their transformed versions with deep features. Unlike facial expressions, audio information possesses richer individual characteristics, making feature selection more difficult. Finding a unified and effective feature selection pattern, along with deep learning architectural methods, remains a crucial task for future research.

The integration of multi-modal features is crucial for future depression detection. In clinical assessments, specialized doctors evaluate the subjects from various perspectives. Similarly, ADE based on deep learning should mimic this approach by extracting and merging features from multiple perspectives and modalities. In particular, methods for feature fusion should be carefully designed by considering common tendencies, temporal synchronicity, and the dynamic nature of modalities. With ongoing enhancements in computational power, ADE methods based on large models will continue to be proposed.

However, data collection and availability remain significant limitations for the development of ADE. First, owing to privacy policies and research ethics, existing open-source datasets are scarce. Second, multi-modal data are rarely used in open-source datasets. While the DAIC-WOZ dataset provides transcripts, audio, and desensitized video information, datasets that offer other features potentially relevant to depression detection are lacking. Third, most of the current research on AI-based depression diagnosis and treatment has a relatively small sample size, making it challenging to accurately reflect the characteristics of the overall population with depression. Fourth, data collection by the different research groups did not follow a unified standard.

In practical applications, deep learning-based ADE methods are still in the early stages of development. Nemesure *et al*[41] assessed the mental well-being of student populations by combining electronic health records with machine learning methods. Aguilera *et al*[42] developed applications and applied them to primary care. We believe that the interpretability of deep learning is a major limitation in its application. Future research should focus on two directions to enhance model credibility. The first is the construction of knowledge-guided ADE models. The research framework proposed by Zhou *et al*[43] is a novel research direction. The second is the incorporation of relevant analyses for model interpretability. Researchers can analyze the operating mechanism of a model using techniques such as visualization and feature capture.

In summary, researchers should focus on ADE based on bodily expressions. Additionally, unified and effective methods for audio feature extraction should continue to be explored. When constructing ADE models, special attention should be paid to the interpretability of the models. We hope that future research will introduce new perspectives and methods to address this aspect. Regarding data collection, research groups should consider publicly sharing their research paradigms, psychological effect evaluations, and desensitization data to help considerably advance the construction of large models and research progress in ADE.

CONCLUSION

In this paper, we provided an overview of prominent audio- and video-based ADE models in recent years, covering the aspects of audio, video, and fusion. An analysis of the relevant research revealed a lack of exploration of the body expressions of individuals with depression. We encourage researchers to delve further into audio feature extraction. In addition, we believe that the construction of large models is crucial for future research. We hope that researchers will develop outstanding ADE models in the future.

90160_Auto_Edited .docx

ORIGINALITY REPORT

2%

SIMILARITY INDEX

PRIMARY SOURCES

1	Jinlong Li, Zhenyu Liu, Zhijie Ding, Gangping Wang. "A novel study for MDD detection through task-elicited facial cues", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018 Crossref	20 words — < 1%
2	irpj.euclid.int Internet	18 words — < 1%
3	www.researchgate.net Internet	16 words — < 1%
4	www.tandfonline.com Internet	13 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS